TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẠI HỌC QUỐC GIA HÀ NỘI

Tần Lê Nghĩa

Nghiên cứu một số phương pháp nâng cao tính tin cậy cho mô hình phân tích sắc thái của bình luận

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công Nghệ Thông Tin

HÀ NỘI – 2022

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẠI HỌC QUỐC GIA HÀ NỘI

Tần Lê Nghĩa

Nghiên cứu một số phương pháp nâng cao tính tin cậy cho mô hình phân tích sắc thái của bình luận

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công Nghệ Thông Tin

Cán bộ hướng dẫn: TS. Lê Đức Trọng,

TS. Hoàng Tuấn Anh

HÀ NỘI – 2022

VIETNAM NATIONAL UNIVERSITY, HANOI

UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Tan Le Nghia

Research on some methods to improve the reliability of comment sentiment analysis model

BACHELOR'S THESIS

Major: Information Technology

Supervisor: Dr. Le Duc Trong,

Dr. Hoang Tuan Anh

LỜI CAM ĐOAN

Em xin cam đoan luận văn tốt nghiệp "Nghiên cứu một số phương pháp nâng cao tính tin cậy cho mô hình phân tích sắc thái của bình luận" là công trình nghiên cứu của em, với sự hướng dẫn của TS Lê Đức Trọng và TS Hoàng Tuấn Anh.

Các số liệu, kết quả được sử dụng trong luận văn là thực sự đạt được đảm bảo tính chính xác trung thực và chưa từng được công bố trong bất kỳ công trình nào khác. Việc sử dụng các thuật ngữ, kết quả của các nghiên cứu khác đều được trích dẫn và để tên tác giả cùng nguồn một cách nghiêm túc và công khai.

Em đã hoàn thành các yêu cầu của nhà trường để được phép tham gia bảo vệ khoá luận. Nếu phát hiện bất kỳ sự gian lận nào, em xin chịu trách nhiệm trước Hội đồng, cũng như kết quả luận văn của mình.

Hà Nội, tháng 4 năm 2022

Sinh viên

Tần Lê Nghĩa

LỜI CẢM ƠN

Lời đầu tiên, em xin được bày tỏ lòng biết ơn đến Ban Giám hiệu trường Đại học Công Nghệ - Đại học Quốc gia Hà Nội, Ban Lãnh đạo nhà trường đã tạo điều kiện cho em có một môi trường học tập, nghiên cứu với chất lượng tốt nhất có thể.

Tiếp theo, em xin chân thành cảm ơn giảng viên hướng dẫn - TS Lê Đức Trọng và TS Hoàng Tuấn Anh về những lời khuyên vô giá và sự chỉ bảo tận tâm của thầy trong thời gian thực tập kỳ hè, và thời gian làm đồ án tốt nghiệp vừa qua. Sự hướng dẫn, hỗ trợ và động viên tận tình của Thầy đã giúp em trong việc nghiên cứu và hoàn thành khoá luận này mặc dù có những thời điểm khó khăn của dịch bệnh ảnh hưởng đến tiến độ của đồ án.

Cuối cùng, em xin bày tỏ tình cảm tới gia đình, người thân và bạn bè đã luôn yêu thương, chia sẻ và giúp đỡ em trong suốt 4 năm qua.

Hà Nội, tháng 4 năm 2022

MỤC LỤC

TÓM TẮT

Ngày nay, cùng với sự phát triển mạnh mẽ và vượt bậc của Internet cùng với sự đa dạng các thiết bị di động, người dùng đang có xu hướng đăng các hình ảnh và bình luận để thể hiện cảm nghĩ, ý kiến của họ về một trải nghiệm nào đó. Vì vậy những trang mạng hiện nay đang được phát triển cho phép người dùng có thể chia sẻ những đánh giá và nhận xét của mình về dịch vụ và sản phẩm của doanh nghiệp, tổ chức. Đặc biệt trong lĩnh vực nhà hàng, người dùng thường tiến hành đặt mua đồ ăn, thức uống hay chọn nhà hàng cho các buổi tiệc thường có xu hướng quan tâm và bị tác động phần nào bởi những bình luận và hình ảnh của những khách hàng trước đó đăng lên. Bên cạnh đó, các doanh nghiệp, nhà hàng cũng tiến hành thu thập các phản hồi của khách hàng về món ăn và dịch vụ của mình để có thể hiểu được khách hàng của mình hơn và đưa ra những chiến lược đúng đắn.

Hiện nay, những thông tin mà các hệ thống được sử dụng để phân tích các phản hồi của người dùng đang chủ yếu tập trung vào dữ liệu văn bản. Tuy nhiên việc chỉ sử dụng dữ liệu văn bản có thể làm cho mô hình phân tích sắc thái bình luận của người dùng không được chính xác. Bên cạnh đó với sự phát triển của các thiết bị di động và Internet dẫn tới việc người dùng đang dần sử dụng những hình ảnh mà mình chụp được để thể hiện những trải nghiệm và cảm xúc của mình. Vì vậy việc kết hợp hình ảnh và bình luận sẽ giúp cho việc phân tích những đánh giá của người dùng được tin cậy và đúng đắn hơn.

Khóa luận với đề tài "Nghiên cứu một số phương pháp nâng cao tính tin cậy cho mô hình phân tích sắc thái của bình luận" tập trung nghiên cứu vào các phương pháp để tăng tính tin cậy cho mô hình phân tích sắc thái bình luận người dùng dành cho tiếng Việt bằng việc đề xuất các mô hình học sâu đa phương thức dựa trên mô hình PhoBert, LSTM, Inception-V3 và Attention. Mô hình được thử nghiệm trên tập dữ liệu thu thập trên trang ẩm thực Foody.

Từ khóa: *phân tích sắc thái, đa phương thức, học sâu, PhoBert, Inception-v3, LSTM, Attention…*

ABSTRACT

Today, along with the strong and outstanding development of the Internet along with the diversity of mobile devices, users are tending to post pictures and comments to express their thoughts and opinions about the Internet. some experience. Therefore, websites are currently being developed that allow users to share their evaluations and comments about the services and products of businesses and organizations. Especially in the restaurant industry, users who often order food and drinks or choose a restaurant for parties tend to be interested and partly influenced by the comments and photos of other guests. previous item posted. Besides, businesses and restaurants also collect customer feedback about their dishes and services to be able to understand their customers better and come up with the right strategies.

Currently, the information that systems are used to analyze user feedback is mainly focused on text data. However, using only text data can make the user comment nuance analysis model inaccurate. Besides, with the development of mobile devices and the Internet, users are gradually using the images they take to express their experiences and emotions. Therefore, the combination of images and comments will make the analysis of user reviews more reliable and correct.

Thesis with the topic "Research on some methods to improve the reliability of the comment sentiment analysis model" focuses on research methods to increase the reliability of the comment nuance analysis model. Vietnamese by suggesting multimodal deep learning models based on PhoBert, LSTM, Inception-V3 and Attention models. The model is tested on the dataset collected on the food site Foody.

Keywords: *sentiment analysis, Multimodal, deep learning, PhoBert, Inception-v3, LSTM, Attention…*

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

DANH MỤC VIẾT TẮT

STT	Từ viết tắt	Cụm từ đầy đủ	Giải thích
1	ANN	Artificial neural network	Mạng nơ-ron nhân tạo
2	LSTM	Long short-term memory	Bộ nhớ dài-ngắn hạn
3	RNN	Recurrent neural netwwork	Mạng nơ-ron hồi quy
4	Seq2seq	Sequence to sequence	Mô hình chuỗi sang chuỗi
5	CNN	Convolutional neural network	Mạng nơ ron tích chập
6	SVM	Support vector machine	Máy vector hỗ trợ
7	RSA-ALF	Reliable sentiment analysis - Attention late fusion	Mô hình phân tích sắc thái bình luận tin cậy – chú ý kết hợp sau
8	RSA-AEF	Reliable sentiment analysis - Attention early fusion	Mô hình phân tích sắc thái bình luận tin cậy – chú ý kết hợp sớm
9	RSA-SM	Reliable sentiment analysis – Stack Max	Mô hình phân tích sắc thái bình luận tin cậy – xếp chồng lớn nhất

Chương 1. Đặt vấn đề

Hiện nay, với sự phát triển mạnh mẽ của Internet, các thiết bị di động và các nền tảng dịch vụ trực tuyến trên Internet, đặc biệt là các trang dịch vụ trong lĩnh vực ẩm thực thì việc phân tích sắc thái bình luận người dùng là một bài toán vô cùng cần thiết hiện nay. Phân tích sắc thái bình luận người dùng sẽ giúp cho các doanh nghiệp, nhà hàng có được cái nhìn tổng quan về các phản hồi của khách hàng về món ăn và dịch vụ của mình để có thể thấu hiểu được khách hàng của mình, biết được các ưu nhược điểm để đưa ra những chiến lược đúng đắn. Từ đó việc tự động phân tích sắc thái bình luận người dùng tự động sẽ giúp các trang thương mại điện tử đặc biệt là về dịch vụ đặt đồ ăn trực tuyến giảm thời gian, chi phí cũng như nhân công cho tác vụ này. Chương 1 sẽ trình bày về vấn đề cần giải quyết, các phương pháp phân tích sắc thái bình luận, hướng tiếp cận và cấu trúc của khóa luận.

1.1. Động lực nghiên cứu

Trong cuộc sống xã hội hiện đại ngày nay, khi mà Internet kết nối toàn cầu và các nền tảng dịch vụ trên nó ngày càng được phát triển mạnh mẽ, đặc biệt chúng ta chứng kiến sử trỗi dậy mạnh mẽ của thị trường giao đồ ăn trực tuyến khi mà các ứng dụng ngày một hoàn thiện hơn. Mặt khác, người dùng thường để lại các đánh giá về đồ ăn của mình trên các trang thương mại điện tử đặt hàng trực tuyến điển hình như Foody, Now…. Ý kiến của khách hàng dù tiêu cực hay tích cực sẽ giúp cho doanh nghiệp, nhà hàng có những phân tích để cải thiện chất lượng dịch vụ của mình.

Để chinh phục khách hàng thì không thể không tìm hiểu về nhu cầu của họ. Tuy nhiên, vấn đề làm sao doanh nghiệp có thể biết được khách hàng đang hài lòng và không hài lòng hay thương hiệu đang được ưa chuộng là gì. Để có thể làm được điều đó các công ty, doanh nghiệp cần phát triển các hệ thống có khả năng phân tích sắc thái bình luận người dùng của mình, phân loại xem bình luận nào là tích cực hay tiêu cực.

Thông thường các hệ thống phân tích sắc thái bình luận các phản hồi của người dùng đang chủ yếu tập trung vào dữ liệu văn bản. Tuy nhiên việc chỉ sử dụng dữ liệu văn bản có thể làm cho mô hình phân tích sắc thái bình luận của người dùng không được chính xác mà chúng ta có thể khai thác những thông tin khác như những hình ảnh mà

người dùng đăng lên được đi kèm với bình luận. Những hình ảnh sẽ giúp phản ánh phần nào chất lượng của món ăn dịch vụ mà người dùng trải nghiệm từ đó hỗ trợ đánh giá thái độ của người dùng.

Ví dụ trong bảng 1 cho chúng ta thấy sự tác động của dữ liệu văn bản như các từ khoá mang ý nghĩa tích cực trong toàn thể câu mang ý nghĩa tiêu cực làm cho mô hình đơn phương thức văn bản (text-only) cụ thể là mô hình PhoBERT+LSTM tuy đã được huấn luyện với độ chính xác (accuracy) đạt 94,22% và macro-f1 đạt 93,16% bị phân loại không chính xác sang nhãn Positive. Tuy nhiên ta có thể thấy hình ảnh chụp không được chỉn chu phần nào giúp ta nhận biết được thái độ không hài lòng của người dùng.

Ảnh người dùng chụp:

Văn bản: cả quán 5 nhân_viên lác_đác vài khách mà đợi hơn 20 mới được 1 cốc cafe phân_nửa các bạn nhân_viên ưu_ái lấy hẳn ống hút to cho đồ nóng tí bỏng gọi bạc sỉu nóng thì lấy nhầm thành nước socola pha loãng mình gọi đúng thế vì ngoại_trừ màu nâu thì là nước_lọc có mùi socola chê xong đến khen nhé quán có món cafe chuối khá lạ hương_vị chuối khá thơm và cafe khá đậm sánh dù hơi ngọt giá_như các bạn phục_vụ nhanh hơn cốc đầy_đặn hơn giá tăng thêm chút cũng đc thì mình nhất_định sẽ trung_thành vs quán

Nhãn: Negative	**PhoBERT+LSTM dự đoán:** Positive

Bảng 1: Ví dụ về vấn đề gặp phải mô hình đơn phương thức

Theo nghiên cứu[1] của Microsoft, mô hình là một hộp đen phức tạp, biến đổi và thường dẫn tới những quyết định không đáng tin cậy so với kì vọng khi triển khai, vì vậy một mô hình cần đảm bảo được ba mục tiêu là sự công bằng, ổn định và có thể giải thích được. Cũng như theo bài báo của Suchi Saria [1] về sự an toàn và tin cậy của mô hình học máy một trong những giải pháp để đảm bảo tính tin cậy cho mô hình học máy đó là tăng thêm thông tin cho mô hình. Những nghiên cứu trên đã cho em những suy nghĩ về những giải pháp cho bài toán mà khoá luận đang cần giải quyết đó là việc kết hợp thêm thông tin ảnh cho mô hình phân tích sắc thái.

Chính những điều trên là động lực để thôi thúc em tìm hiểu một số phương pháp nâng cao tính tin cậy cho mô hình phân tích sắc thái bình luận để việc phân tích ý kiến của người dùng được chính xác tin cậy hơn.

1.2. Một số hướng tiếp cận hiện có

Từ những năm 2000 cho đến nay, phân tích cảm xúc đã và đang thu hút được các nhà nghiên cứu quan tâm, phát triển và đưa vào áp dụng thực tế. Khái niệm phân tích cảm xúc (sentiment analysis) xuất hiện lần đầu trong công trình của Nasukawa và Yi [2]. Tuy nhiên, nghiên cứu được xem là nền móng cho sự phát triển của lĩnh vực phân tích ý kiến sau này là nghiên cứu của Pang và các cộng sự [3] :

Công trình [3] đã tiến hành nghiên cứu về phân tích ý kiến từ các phản hồi của người dùng đối với miền dữ liệu điện ảnh (movie domain) với hai lớp được quan tâm đến là tích cực và tiêu cực. Ba phương pháp máy học (Maximum entropy classification ,naive bayesvà support vector machine) được sử dụng để giải quyết bài toán phân loại bình luận trong nghiên cứu này.

Bên cạnh những thành tựu nghiên cứu trên thế giới, bài toán phân tích cảm xúc cũng thu hút được sự quan tâm của cộng đồng nghiên cứu trong nước trên đa dạng các

[1] https://www.microsoft.com/en-us/research/group/reliable-machine-learning/publications/

miền dữ liệu khác nhau như nhà hàng, khách sạn, giáo dục,. . . v.v. Theo như em tìm hiểu, một trong những công trình nghiên cứu đầu tiên về phân tích ý kiến trên tiếng Việt được thực hiện bởi Kieu & Pham [4] trên cấp độ câu văn, tiến hành thực nghiệm đánh giá trên bộ ngữ liệu về miền dữ liệu máy tính đạt được độ đo F1 là 62.84%.

Gần đây, Lắc và cộng sự [5] đã sử dụng sự kết hợp của Bidirectional Long-Short Term Memory với CNN (BiLSTM-CNN) trên hai bộ dữ liệu, đánh giá của khách hàng trên các nền tảng thương mại điện tử, cụ thể là Tiki và bộ dữ liệu về các phản hồi của các bạn sinh viên trong quá trình học tập (Vietnamese Student's Feedback Corpus). Và kết quả công trình nghiên cứu mang lại, đạt được 93.55% với thang đo F1 trên bộ dữ liệu VSFC và 84.14% trên bộ dữ liệu Tiki, mang lại kết quả tốt tăng 2.36% - 8.55% so với các công trình nghiên cứu ở thời điểm trước.

1.3. Phát biểu bài toán

Phân tích sắc thái (sentiment analysis) là quá trình xác định, phân loại các văn bản thành các thái độ, tình cảm đối với một vấn đề nào đó.

Bài toán phân tích cảm xúc hiện nay có thể phân thành 3 cấp độ câu văn (sentence-level), văn bản (document-level), và khía cạnh (aspect-level). Ở cấp độ câu văn, mục tiêu của bài toán là từ một câu văn phân thành các lớp nhãn. Cấp độ văn bản được dùng để xác định mức độ cảm xúc của một đoạn văn (gồm hai hay nhiều câu văn). Và cấp độ khía cạnh được dùng để xác định mức độ cảm xúc cho mỗi khía cạnh của thực thể đề cập trong một văn bản. Trong phạm vi của đồ án, giới hạn nghiên cứu nhóm sẽ chỉ nằm ở cấp độ câu văn.

Đối với bài toán phân tích sắc thái bình luận, các nghiên cứu đã tiến hành/thử nghiệm nhiều phương án tiếp cận khác nhau, từ các phương pháp truyền thống sử dụng SVM, NaiveBayes,..cho đến các mạng học sâu CNN, BiLSTM...Tuy nhiên, các công trình nghiên cứu vẫn đang chỉ tập trung sử dụng thông tin từ dữ liệu văn bản để phân tích sắc thái bình luận, điều này dẫn tới sự thiếu tin cậy và không chính xác nếu như dữ liệu văn bản chứa những thông tin làm ảnh hưởng đến chất lượng phân loại của mô hình. Từ vấn đề đó, đồ án đề xuất các mô hình đa phương thức kết hợp dữ liệu hình ảnh và văn bản nhằm nâng cao tính tin cậy và độ chính xác.

Đầu vào của bài toán là câu bình luận người dùng và tập ảnh đi kèm với câu bình luận đó, đầu ra sẽ là phân loại hai nhãn Positive và Negative. Sử dụng cơ chế attention để tính trọng số cho các ảnh nhằm đưa ra vector ngữ cảnh cho ảnh sau đó kết hợp với vector văn bản để tạo ra một vector mới giúp cho mô hình phân loại sắc thái bình luận.

1.4. Bố cục của khóa luận

Khóa luận gồm 4 phần chính tương ứng với 4 chương của khoá luận:

- Chương 1: Đặt vấn đề và các phương pháp hiện có cho bài toán, hướng tiếp cận của khóa luận
- Chương 2: Giới thiệu các khái niệm, cơ sở lý thuyết của bài toán phân tích sắc thái bình luận
- Chương 3: Xây dựng mô hình phân tích sắc thái bình luận dựa trên mạng PhoBert, Inception-V3, LSTM và attention và đánh giá kết quả thực nghiệm trên dữ liệu Foody
- Chương 4: Kết luận và hướng phát triển

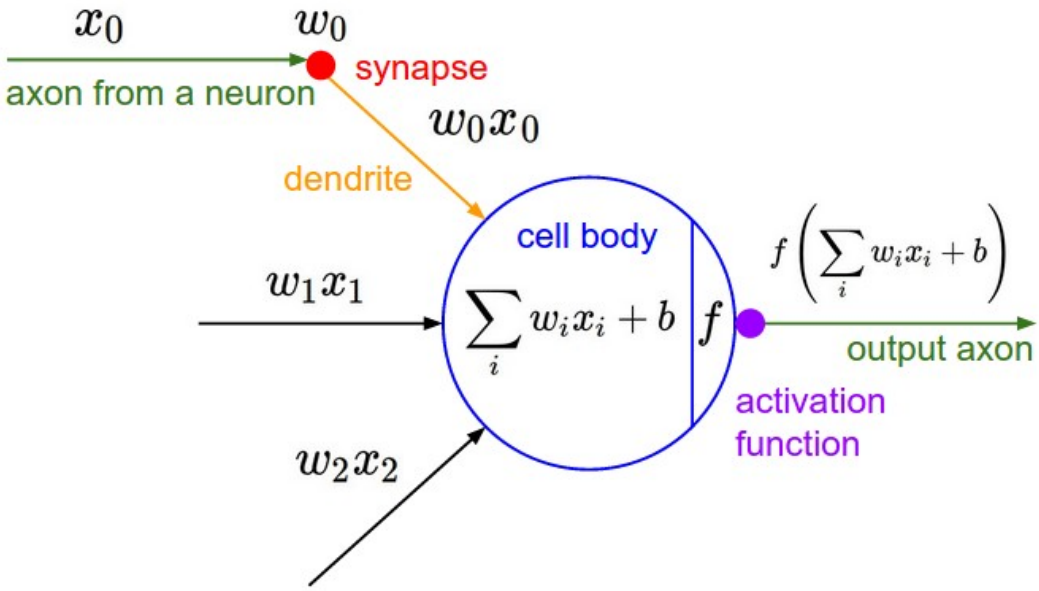
Chương 2. Cơ sở lý thuyết

2.1. Giới thiệu về mạng nơ-ron nhân tạo (artificial neural network)

Mạng nơ ron nhân tạo là mô hình được lấy cảm hứng từ hệ thần kinh não bộ của con người. Mạng nơ ron nhân tạo bao gồm một tập các nút hoạt động như các nơ-ron thần kinh trong hệ thần kinh của con người và một tập các cạnh nối các nút với nhau như các dây thần kinh để truyền thông tin từ nút thần kinh này sang nút thần kinh khác.

2.1.1. Nơ ron nhân tạo trong mạng nơ ron nhân tạo

Nơ ron nhân tạo là thành phần cơ bản để cấu tạo nên mạng nơ ron nhân tạo, một nơ-ron thường được gọi là một nút mạng (unit), đầu vào của một nút sẽ là thông tin đầu vào từ bên ngoài mạng hoặc đầu ra của các nút phía trước. Mỗi nút sẽ có một tập các tham số để thực hiện tính toán và biến đổi bao gồm trọng số weight và có thể có thêm bias. Mỗi một nút sau khi nhận đầu vào sẽ sử dụng trọng số weight và bias (nếu có tuỳ vào trưởng hợp) để thực hiện tính toán cho đầu ra của nút, việc tính toán của một nút giống như một hàm số nhận giá trị đầu vào, sau đó thực hiện các biến đổi từ đầu vào đó để tính toán ra đầu ra như mong muốn. Hình dưới đây minh họa một nơ ron trong mạng nơ ron nhân tạo hình ảnh được tham khảo từ [6]:



Hình 2-1: Mô phỏng một nơ-ron nhân tạo trong mạng nơ-ron nhân tạo

Trong hình trên nút sẽ nhận đầu vào là x_1 và x_2, sử dụng trọng số w_1, w_2, b để tính ra đầu ra:

$$y = f(w_1 x_1 + w_2 x_2 + b)$$

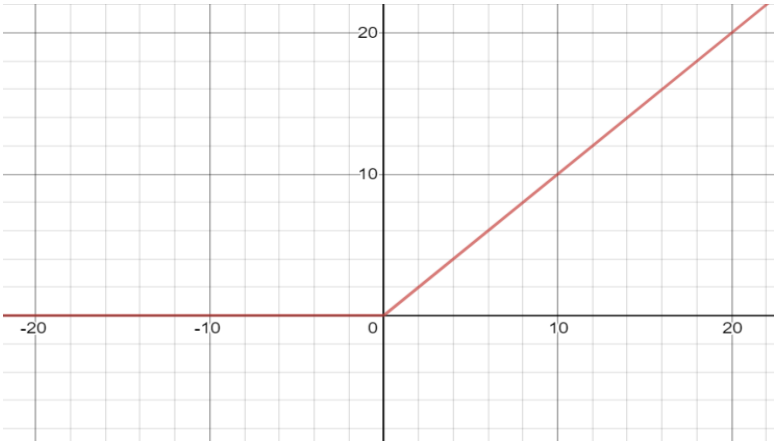
Trong đó b là bias được tính bằng $b = w_0 x_0$

2.1.2. Hàm kích hoạt (activation functions)

Hàm kích hoạt là một thành phần trong các mạng nơ ron nhân tạo. Mỗi một lớp trong mạng nơ ron nhân tạo đều cần một hàm kích hoạt dùng để biến đổi từ tuyến tính thành phi tuyến tính các thông tin từ lớp trước nó. Nếu không sử dụng hàm kích hoạt thì dù cho có sử dụng bao nhiêu lớp ẩn đi nữa thì bản chất mạng nơ ron đó vẫn chỉ thực hiện được biến đổi tuyến tính.

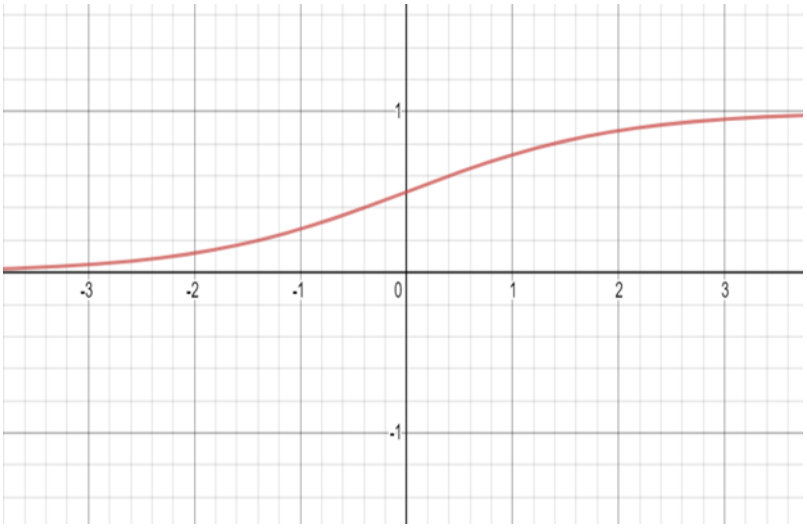
Một số hàm kích hoạt thông dụng:

- ReLU: $f(x) = \max(0, x)$



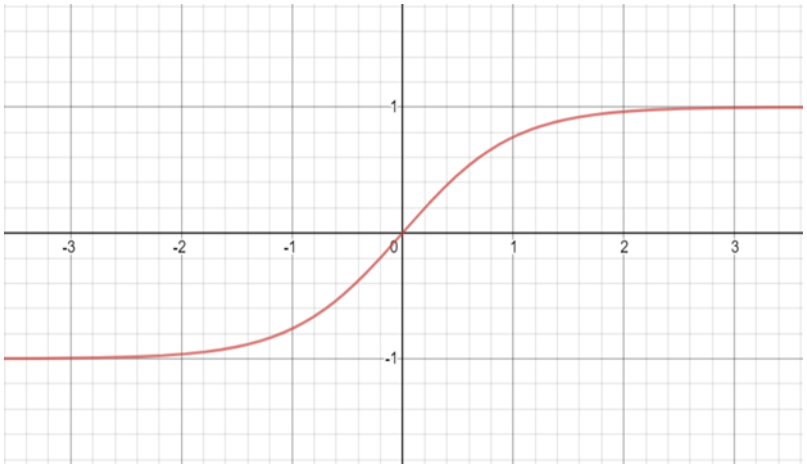
Hình 2-2: Đồ thị hàm ReLU

- Sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$



Hình 2-3: Đồ thị hàm Sigmoid

- Tanh: $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$

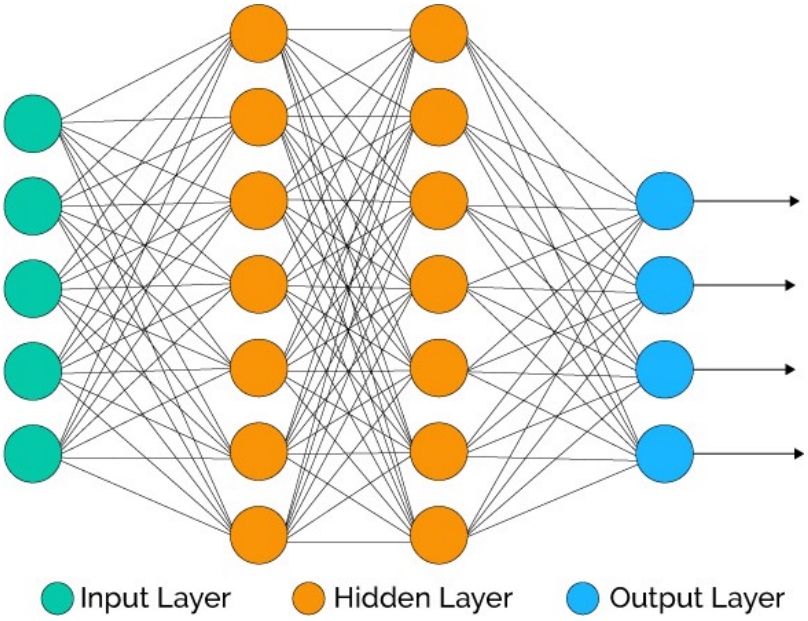


Hình 2-4: Đồ thị hàm tanh

2.1.3. Mạng truyền thẳng

Mạng nơ ron truyền thẳng là mạng nơ ron nhân tạo có cấu trúc đơn giản nhất. Nó bao gồm nhiều nút (node) trong một lớp (layer), các nút trong một lớp có các liên kết nối với các nốt ở các lớp kề tiếp đó, và các nốt trong cùng một lớp không kết nối với nhau. Dưới đây là minh họa của một mạng truyền thẳng.

Hình ảnh được tham khảo từ [6] minh họa một mạng truyền thẳng thường bao gồm 3 thành phần lớp:



Hình 2-5: Mạng truyền thẳng đơn giản

- Lớp đầu vào (input layer): Lớp đầu vào sẽ nhận thông tin các đặc trưng của đầu vào và truyền nó đến các lớp ẩn phía sau, các nốt trong lớp đầu vào sẽ không thực hiện bất kì một tính toán nào.
- Các lớp ẩn (hidden layers): Theo sau lớp đầu vào, sẽ là một hoặc nhiều lớp ẩn, các lớp ẩn sẽ nhận thông tin từ lớp ngay phía trước nó làm đầu vào và thực hiện một số biến đổi và truyền tiếp đầu ra của nó cho lớp ngay phía sau nó.
- Lớp đầu ra (output layer): Đây là lớp cuối cùng trong mạng truyền thẳng. Lớp này sẽ nhận thông tin từ lớp ẩn cuối cùng trong các lớp ẩn, thực hiện tính toán biến đổi để cho ra đầu ra của bài toán.

Trong mạng truyền thẳng thông tin chỉ di chuyển một chiều từ lớp đầu vào qua các lớp ẩn đến lớp đầu ra để tạo ra đầu ra cho bài toán. Không có đường truyền ngược nào từ lớp sau về các lớp phía trước trong mạng truyền thẳng.

2.2. Các phương pháp mô hình hoá thông tin ngữ nghĩa văn bản

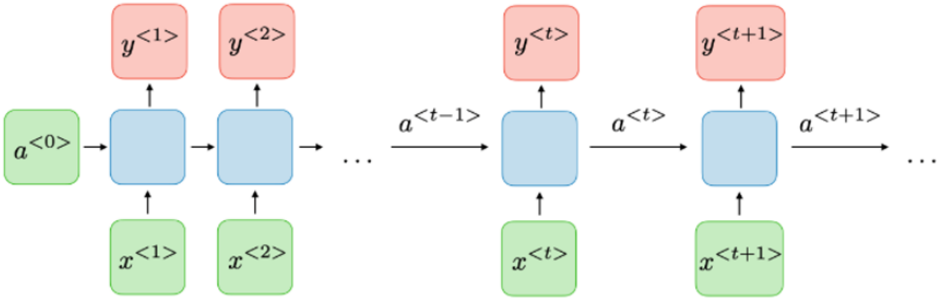
2.2.1. Mạng nơ-ron hồi quy (RNN)

Mạng nơ-ron hồi quy (RNN) là một loại đặc biệt của mạng nơ-ron nhân tạo được điều chỉnh để phù hợp cho dữ liệu chuỗi thời gian hoặc dữ liệu dạng chuỗi. Các mạng

nơ-ron chuyển tiếp thông thường chỉ dành cho các điểm dữ liệu, các điểm này độc lập với nhau. Tuy nhiên, nếu chúng ta có dữ liệu theo một trình tự sao cho một điểm dữ liệu phụ thuộc vào điểm dữ liệu trước đó, chúng ta cần sửa đổi mạng nơ-ron để kết hợp sự phụ thuộc giữa các điểm dữ liệu này. Các mạng hồi quy có khái niệm 'bộ nhớ' giúp chúng lưu trữ các trạng thái hoặc thông tin của các đầu vào trước đó để tạo ra đầu ra tiếp theo của chuỗi. Những kiến thức và hình ảnh trong phần 2.2 này được tham khảo từ [7].

2.2.2. Quá trình lan truyền mạng nơ-ron hồi quy (RNN)

Trong mạng nơ ron hồi quy, đầu vào ở một thời điểm t sẽ gồm phần tử thứ t trong chuỗi và trạng thái ẩn (hidden state) thứ t -1, trạng thái ẩn (hidden state) mới sẽ được tính dựa trên trạng thái ẩn ở bước trước và tiếp tục truyền tiếp qua các bước tiếp theo. Chính vì lý do đó mà mạng được gọi là hồi quy, có thể coi trạng thái ẩn (hidden state) đó như một bộ nhớ được sử dụng để lưu lại thông tin từ các bước trước đó. Trên lý thuyết thì mạng hồi quy có thể xử lý các chuỗi với độ dài tùy ý, tuy nhiên trên thực tế nó gặp phải một số hạn chế trong việc xử lý chuỗi dài, vấn đề này sẽ được nói ở phần sau. Dưới đây là minh họa cho mạng hồi quy và quá trình xử lý chuỗi tuần tự của mạng hồi quy:



Hình 2-6: Mạng hồi quy (RNN)

Trong hình trên là quá trình luồng dữ liệu được xử lý trong mạng hồi quy.

- $a^{<i>}$ đại diện cho trạng thái ẩn ở thời điểm i
- $x^{<i>}$ là phần tử thứ i trong chuỗi đầu vào
- $y^{<i>}$ là đầu ra ở thời điểm thứ i
- i=0,1,2……t+1.

Ở thời điểm đầu tiên, trạng thái ẩn được khởi tạo là $a^{<0>}$ (thông thường $a^{<0>}$ thường là véc tơ được khởi tạo ngẫu nhiên hoặc gồm toàn phần tử 0)

Với mỗi thời điểm t, trạng thái ẩn $a^{<t>}$ và đầu ra $y^{<t>}$ sẽ được tính bằng công thức:

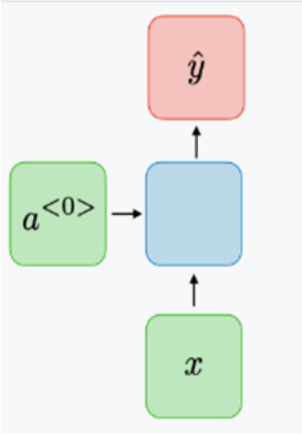
- $a^{<t>} = f(W_{aa}\, a^{<t-1>} + W_{ax}\, x^{<t>} + b_a)$
- $y^{<t>} = g\,(W_{ya}\, a^{<t>} + b_y)$

Trong đó W_{aa}, W_{ax}, W_{ya}, b_a, b_y là các tham số của mạng hồi quy còn f và g là các hàm kích hoạt.

Đối với mạng truyền thẳng truyền thống, mỗi lớp ẩn sẽ có một bộ tham số riêng, còn với mạng hồi quy bộ tham số (W_{aa}, W_{ax}, W_{ya}, b_a, b_y) sẽ được chia sẻ cho tất cả các bước xử lý theo chuỗi thời gian, điều này giúp mạng hồi quy giảm đáng kể số lượng tham số cần học so với mạng truyền thẳng.

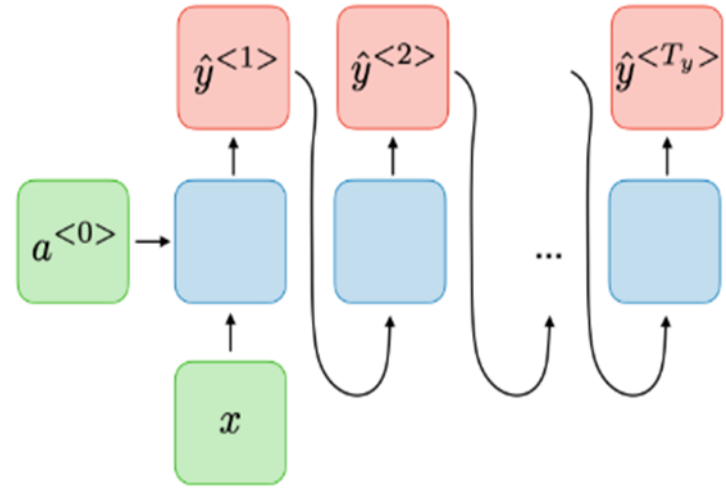
Trong thực tế mạng hồi quy được ứng dụng dưới nhiều dạng, tiêu biểu như một số kiểu sau:

One to one: Mạng RNN kiểu này sẽ bao gồm một phần tử đầu vào và một phần tử đầu ra, đây là kiểu đơn giản nhất.



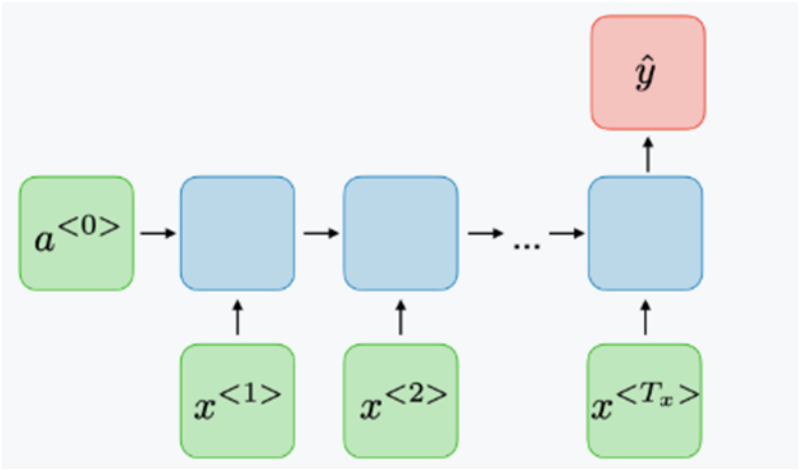
Hình 2-7: Mạng RNN kiểu one-to-one

One to many: Mạng RNN kiểu này có một phần tử đầu vào và có nhiều phần tử đầu ra, kiểu này có thể sử dụng trong các bài toán như mô hình ngôn ngữ



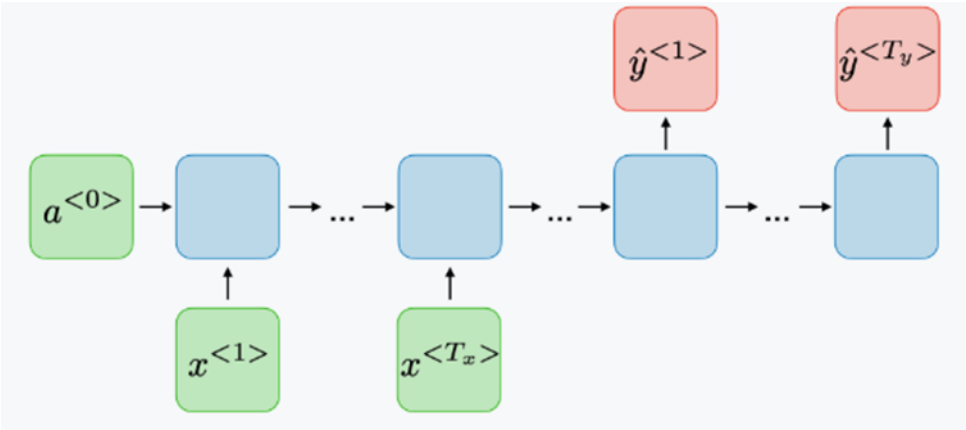
Hình 2-8: Mạng RNN kiểu one-to-many

Many to one: Mạng RNN kiểu này có nhiều phần tử đầu vào và chỉ có một đầu ra, kiểu này được sử dụng trong bài toán phân loại.

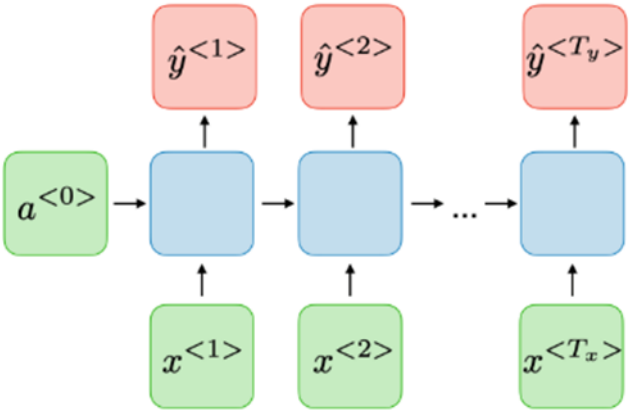


Hình 2-9: Mạng RNN kiểu many-to-one

Many to many: Mạng RNN có nhiều phần tử đầu vào và cũng có nhiều phần tử đầu ra tuy nhiên dạng này còn chia làm hai kiểu nhỏ hơn. Kiểu thứ nhất có số lượng phần đầu vào bằng với số phần tử đầu ra, được ứng dụng trong bài toán nhận diện thực thể tên riêng (named-entity recognition). Kiểu thứ hai có số phần tử đầu vào và số phần tử đầu ra khác nhau, thường được ứng dụng trong bài toán dịch máy (machine translation).



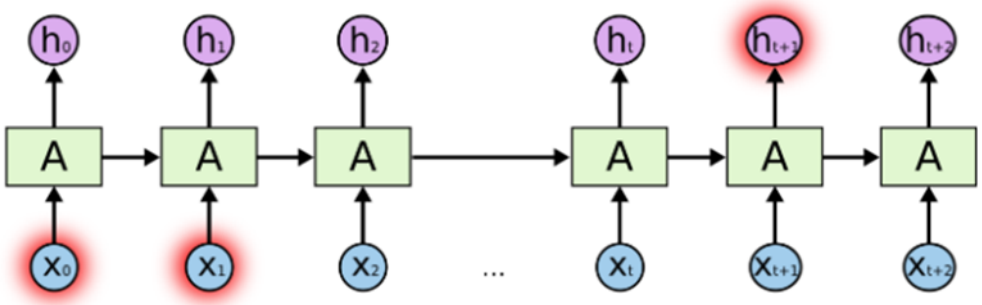
Hình 2-10: Mạng RNN dạng many-to-many với độ dài chuỗi đầu vào và đầu ra bằng nhau



Hình 2-11: Mạng RNN dạng many-to-many với độ dài chuỗi đầu vào và đầu ra khác nhau

2.2.3. Vấn đề phụ thuộc xa trong mạng hồi quy (Long term dependencies)

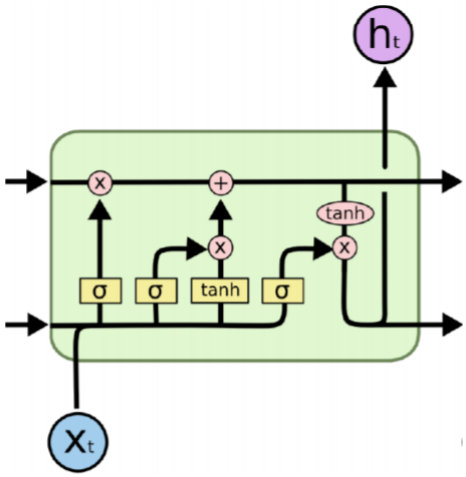
Ưu điểm của mạng hồi quy là lưu trữ các thông tin quá khứ để tính toán đầu ra hiện tại. Tuy nhiên nếu việc tính toán đạo hàm được lặp lại nhiều lần qua nhiều nút sẽ gây nên hiện tượng đạo hàm bị triệt tiêu tiến tới 0 dẫn tới hiện tượng vanishing gradient. Giá trị này trở nên vô cùng nhỏ tại các lớp nơ-ron đầu tiên khiến cho các lớp đó không thể cập nhật trọng số mạng. Ví dụ như khi thực hiện nhiệm vụ đoán từ tiếp theo trong câu "Nó rất thích gặm xương" thì khi đã biết câu "Vàng là một chú chó" thì ta dễ dàng đoán được từ tiếp theo là "xương" dựa trên ngữ cảnh các từ trước đó. Tuy nhiên với đoạn "Tôi là người Sài Gòn. Tôi đang sống ở Hà Nội. Tôi chỉ biết nói tiếng miền Nam" thì để đoán từ "Nam" sẽ cần thông tin từ các trạng thái ở rất xa, tuy nhiên mạng hồi quy khi gặp hiện tượng này sẽ không thể lưu trữ được thông tin đó.



Hình 2-12: Mô tả sự phụ thuộc xa trong RNN

2.2.4. Mạng nơ ron có nhớ LSTM

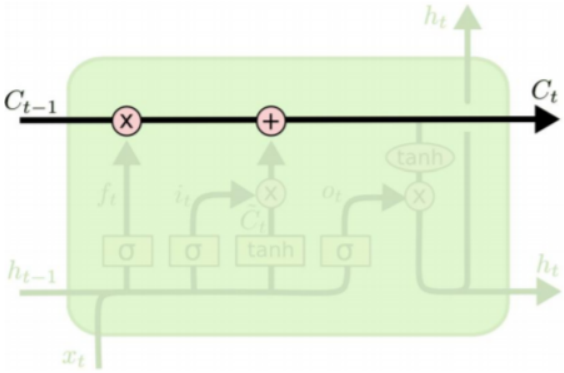
Long short-term memory (LSTM) mạng trí nhớ ngắn hạn định hướng dài hạn đã được giới thiệu bởi Hochreiter và Schmidhuber năm 1997 [8], được sử dụng rộng rãi đến ngày nay và khắc phục được vấn đề tiêu biến (vanishing gradient) cũng như bùng nổ đạo hàm (exploding gradient) trên các mô hình hồi quy RNN. Sự thành công của LSTM là tại mỗi một nút mạng LSTM (LSTM cell) sẽ có thêm một bộ nhớ dài hàn là cell state để lưu trữ thông tin dài hạn được tính toán kết hợp từ 3 loại cổng (gate) gồm cổng quên, cổng vào, cổng ra để quản lí các thông tin tại mỗi nút. Các kiến thức trong phần này được tham khảo từ [9].



Hình 2-13: Kiến trúc LSTM

Trạng thái nhớ của LSTM chính là trạng thái nhớ (cell state), biểu diễn ở đường kẻ ngang trên cùng của Hình 2-14.

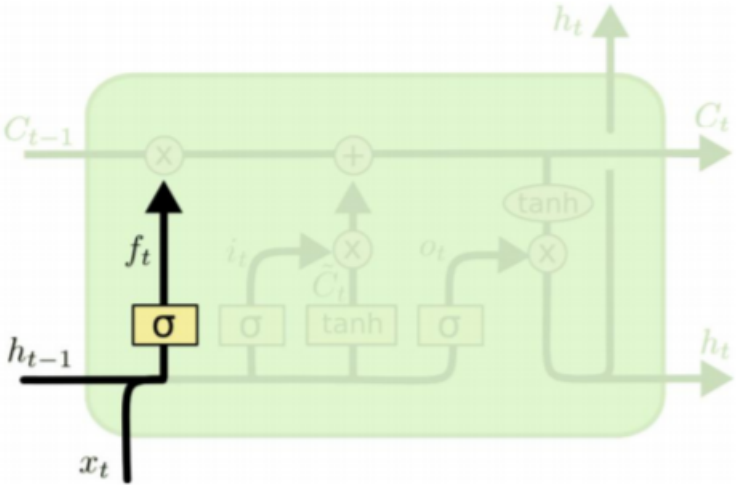
Véc tơ nhớ C_{t-1} được đưa vào một ống nhớ (memory pipe) và có thể xoá hoặc thêm thông tin vào trạng thái nhớ (cell state) được thực hiện và quản lí bởi 3 cổng.



Hình 2-14: Ống nhớ trong khối LSTM

Cụ thể từng bước hoạt động của LSTM [6] như sau:

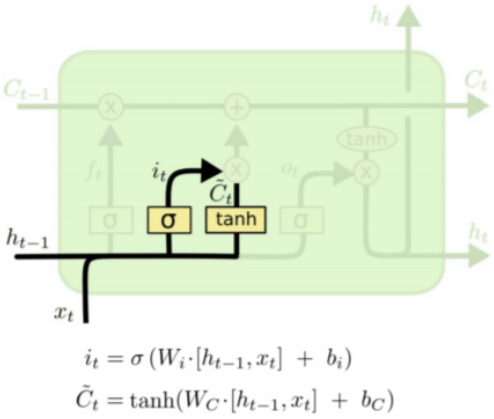
Bước đầu tiên trong LSTM là quyết định thông tin nào sẽ được loại bỏ khỏi cell state. Quá trình quyết định này được thực hiện qua một lớp sigmoid gọi là "forget gate layer" thực hiện. Cổng quên lấy đầu vào là h_{t-1} và x_t và cho đầu ra là một giá trị nằm trong khoảng [0, 1] trong cell state C_{t-1}. Nếu kết quả đầu ra là 1 thể hiện cho việc "giữ lại thông tin", và 0 thể hiện rằng "thông tin sẽ được loại bỏ".



Hình 2-15: Cổng bỏ nhớ của LSTM

Bước tiếp theo sẽ là quyết định thông tin mới nào sẽ được lưu lại tại cell state. Bước này gồm hai phần, đầu tiên h_{t-1} và x_t sẽ được đưa vào tính toán và qua một lớp

sigmoid gọi là “input gate layer” (lớp đầu vào) quyết định giá trị sẽ tích hợp vào trạng thái mới, tiếp theo một lớp tầng ẩn tanh sẽ tạo ra một véc tơ các giá trị mới, $\check{C}_t$, mà có thể được thêm vào cell state hiện tại.

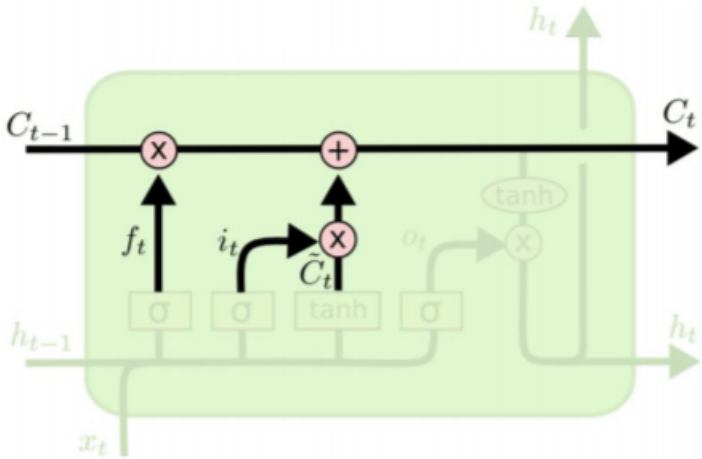


Hình 2-16: LSTM tính toán giá trị lưu tại cell state

Kế tiếp, trạng thái cell state cũ C_{t-1} sẽ được cập nhật sang một trạng thái cell state mới C_t theo công thức:

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

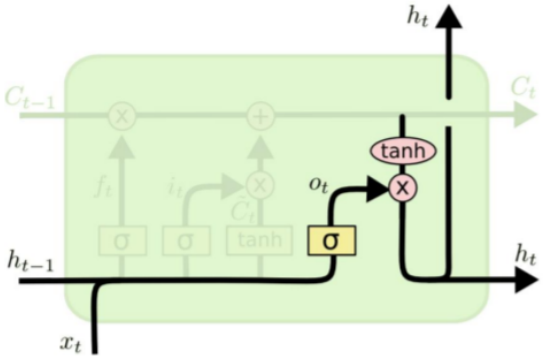
Trạng thái nhớ cũ C_{t-1} được nhân với cổng quên f_t để giữu lại thông tin. Giá trị $i_t * C_t$ thể hiện cho việc cập nhật giá trị cho mỗi cell state. Hình 2-19 minh họa việc cập nhật giá trị cho cell state tại bước này.



Hình 2-17: Cập nhật giá trị Cell State

Bước cuối cùng, khối LSTM quyết định đầu ra của nó dựa trên cell state được minh họa trong hình 2-18. Lớp sigmoid được dùng để tính toán thành phần của cell state sẽ được xuất ra. Sau đó, giá trị cell state được đưa vào hàm tanh (kết quả sẽ thuộc khoảng [-1,1]) và nhân với kết quả đầu ra với cổng ouput, để quyết định cái gì sẽ được khối LSTM xuất ra. Công thức tính toán cho các thành phần của bước này như sau:

$$o_t = \sigma(\mathrm{W}_0[h_{t-1}, x_t] + b_0)$$
$$h_t = o_t * \tanh(C_t)$$

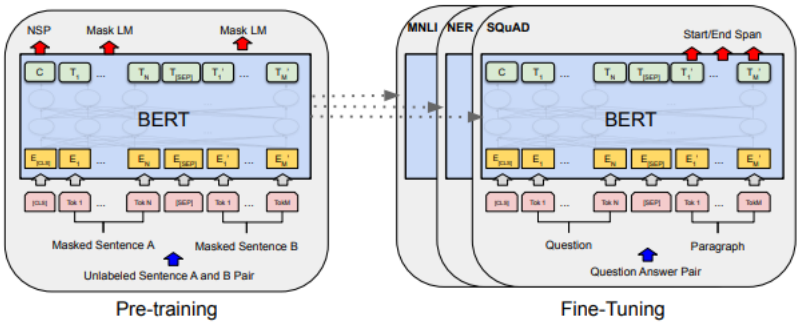


Hình 2-18: Đầu ra của khối LSTM

Trạng thái h_t và cell state C_t hiện tại sẽ tiếp tục làm đầu vào cho LSTM cell tiếp theo.

2.2.5. Mô hình PhoBERT

PhoBERT là mô hình ngôn ngữ đào tạo trước ở cấp độ từ (word-level) đầu tiên cho tiếng việt được giới thiệu bởi VinAI năm 2020 [10]. PhoBERT bao gồm hai phiên bản và có kiến trúc giống với BERT[11]. BERT là viết tắt của Bidirectional Encoder Representations from Transformers, biểu diễn thể hiện mã hóa hai chiều từ Transformer được phát triển bởi Jacob Devlin và cộng sự từ Google vào năm 2018 [11]. $\text{PhoBERT}_{\text{base}}$ với kiến trúc mạng gồm 12 lớp, 768 lớp ẩn, 12 đầu, 110 triệu tham số, và mô hình $\text{PhoBERT}_{\text{large}}$ với kiến trúc mạng gồm 24 lớp, 1024 lớp ẩn, 16 đầu, 340 triệu tham số. BERT có thể học ngữ cảnh của từ theo hai chiều trái phải khác với các mô hình không ngữ cảnh như Word2Vec, Glove hay một chiều như RNN.



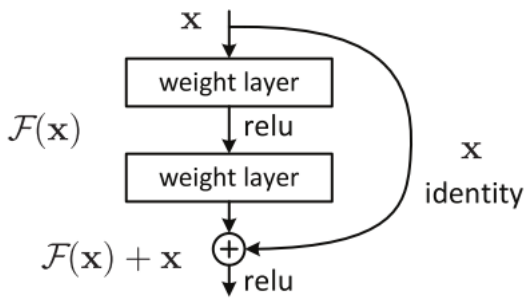
Hình 2-19: Kiến trúc mô hình BERT được trình bày trong

Cách tiếp cận huấn luyện trước của PhoBERT dựa trên RoBERTa là một cách tiếp cận để tối ưu hiệu năng cho mô hình BERT. Mô hình PhoBERT được huấn luyện trên 20GB dữ liệu tiếng Việt (Wikipedia + tin tức).

2.3. Các phương pháp mô hình hoá thông tin ảnh

2.3.1. Mô hình Resnet

Mô hình Resnet viết tắt của residual network được giới thiệu bởi Kaiming He năm 2015 [12]. Ý tưởng của mô hình chính là các kết nối 'tắt' đồng nhất để đi qua một hay nhiều lớp tích chập từ đó giải quyết vấn đề của các mạng nơ ron tích chập về vấn đề vanishing gradient khi mạng có quá nhiều lớp theo chiều sâu. Hình 2-20 thể hiện một khối phần dư của Resnet. Các kiến thức và hình ảnh trong phần này được tham khảo từ [13].



Hình 2-20: Khối kết nối phần dư

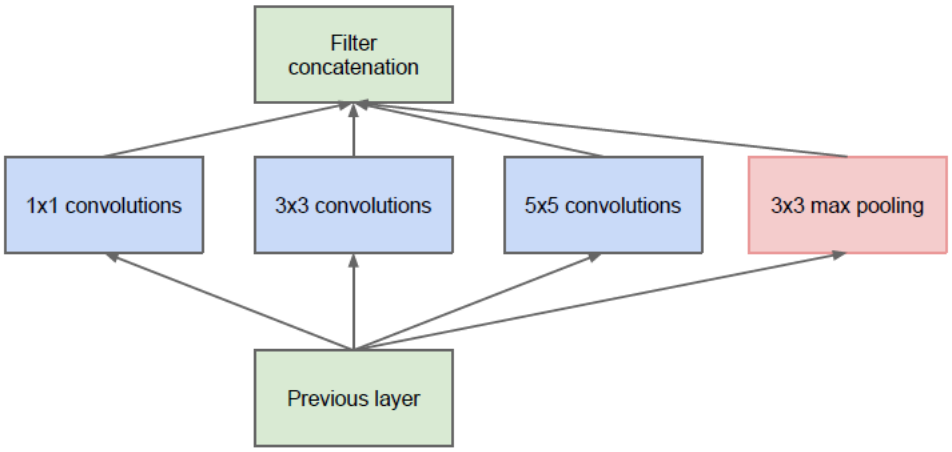
Từ hình 2-20 ta thấy được mũi tên cong đi qua các lớp tích chập đây chính là kết nối 'tắt'. Nói cách khác mạng sẽ bổ sung input X vào đầu ra của layer, hay chính là phép cộng F(x) + x, điều này sẽ tránh cho việc đạo hàm bằng 0, do đạo hàm của x luôn là 1.

Với H(x) = F(x) + x là giá trị dự đoán, F(x) là giá trị nhãn, mô hình sẽ tối ưu sao cho H(X) gần với F(X).

2.3.2. Mô hình Inception

Mô hình Inception được giới thiệu bởi đội ngũ Google năm 2014 [14] được đánh giá là mô hình mới và mang tính hiệu quả cao trong thời điểm bấy giờ. Inception giải quyết vấn đề nếu như có quá nhiều lớp convolution trong mạng theo chiều sâu sẽ dẫn tới hiện tượng overfitting, vì vậy Inception sử dụng ý tưởng sử dụng nhiều lớp lọc trên cùng một bậc phát triển mạng theo chiều rộng thay vì chiều sâu. Các kiến thức và hình ảnh trong phần này được tham khảo từ [15].

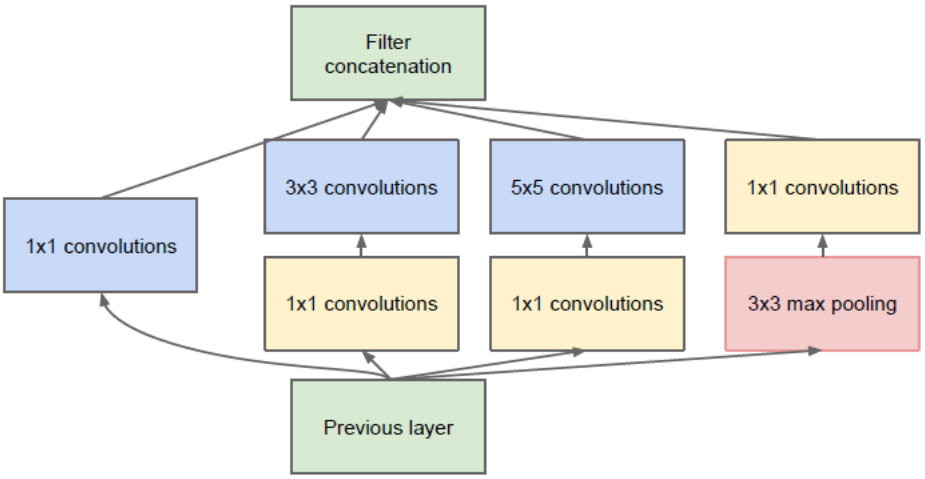
Cấu trúc cơ bản của Inception mô đun được tạo từ 4 filter song song: tích chập 1x1, tích chập 3x3, tích chập 5x5, maxpooling 3x3. Tác dụng lớp tích chập 3x3 và 5x5 được sử dụng giúp cho mạng học được những thành phần không gian của ảnh ở nhiều kích thước hơn. Lớp tích chập 1x1 có tác dụng giúp mô hình học được các đặc trưng theo chiều sâu của ảnh. Lớp maxpooling giúp giảm kích thước ảnh đồng thời giúp mô hình học được những đặc trưng mang tính khái quát hơn. Hình 2-21 mô tả mô đun inception dạng naiveform



Hình 2-21: Mô đun Inception dạng Naiveform

Tuy nhiên dạng naiveform sẽ dẫn tới độ phức tạp về mặt tính toán cho mô hình do lớp tích chập 3x3 và 5x5 sẽ cần tiêu tốn rất nhiều phép tính vì vậy một lớp tích chập 1x1 sẽ được thêm vào trước mỗi bộ lọc tích chập trước giúp giảm số kênh ảnh (bottle-neck)

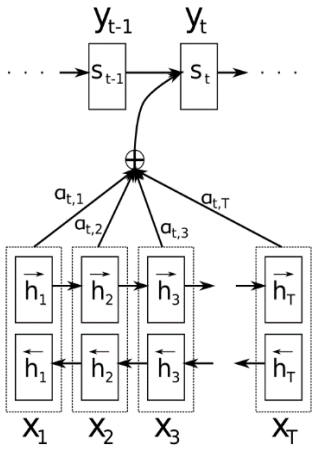
và từ đó tăng tốc độ tính toán cho mô hình. Hình 2-22 mô tả mô đun Inception sau khi áp dụng giảm kích thước



Hình 2-22: Mô đun Inception cùng với giảm kích thước

2.4. Cơ chế Attention

Những thông tin trong phần này dựa trên hai bài báo khoa học và cuốn sách "Deep Learning" của các tác giả Ian Goodfellow, Yoshua Bengio và Aaron Courville [16]. Attention là một khái niệm có đóng góp quan trọng trong cộng đồng học sâu, nó lần đầu được đề xuất bởi Bahdanau et al [17] và sau đó được định nghĩa lại bởi Luong et al [18]. Ban đầu nó được ra đời nhằm nâng cao hiệu quả trong bài toán dịch máy (nerural machine translation) sử dụng kiến trúc encoder-decoder. Với kiến trúc này encoder sẽ xử lý chuỗi đầu vào và tạo ra một véc tơ ngữ cảnh (tức véc tơ trạng thái ẩn cuối cùng) tổng hợp lại thông tin của chuỗi đầu vào. Decoder sẽ nhận véc tơ ngữ cảnh từ encoder và thực hiện đoán từng phần tử của chuỗi đầu ra. Điểm yếu của mô hình này nằm ở chỗ véc tơ ngữ cảnh sẽ có độ dài bị cố định do đó thông tin của chuỗi đầu vào thể hiện trong véc tơ ngữ cảnh sẽ bị giới hạn nếu chuỗi đầu vào có độ dài lớn. Để khắc phục vấn đề này attention sẽ lấy tất cả các véc tơ thể hiện trạng thái ẩn khi xử lý từng phần tử của chuỗi đầu vào, sau đó tính mức độ quan trọng của mỗi véc tơ trạng thái ẩn đó trong việc dự đoán từng phần tử của chuỗi đầu ra.



Hình 2-23: Minh họa cơ chế Attention

Thay vì dùng duy nhất véc tơ trạng thái ẩn cuối cùng của encoder làm véc tơ ngữ cảnh như mô hình encoder-decoder truyền thống, véc tơ ngữ cảnh có sử dụng attention được tính như sau:

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j$$

Trong đó $a_{i\,j}$ là trọng số của trạng thái ẩn h_j trong khi tính c_i, $a_{i\,j}$ được tính bằng công thức dưới đây:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

eij là điểm trọng số đánh giá mức độ phụ thuộc mạnh hay yếu giữa trạng thái ẩn sj của chuỗi đầu ra so với trạng thái ẩn hi ở chuỗi đầu vào. Dưới đây là một số cách tính eij phổ biến.

Name	Alignment score function
Content-base attention	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$
Additive(*)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \boldsymbol{s}_t)$ Note: This simplifies the softmax alignment to only depend on the target position.
General	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix in the attention layer.
Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$
Scaled Dot-Product(^)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.

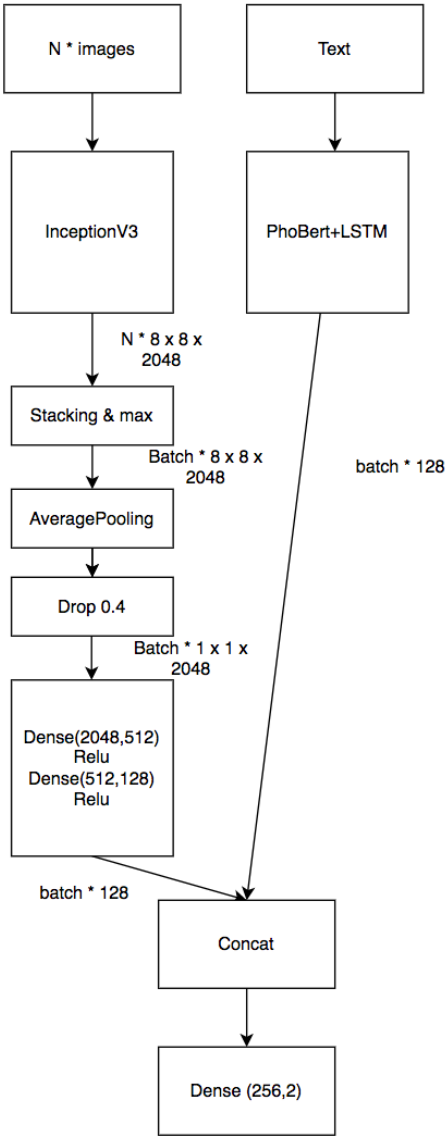
Hình 2-24: Một số cách tính điểm trọng số trong attention

Chương 3. Các mô hình đề xuất và kết quả thực nghiệm

Chương 3 sẽ trình bày các mô hình đề xuất và kết quả thử nghiệm trên bộ dữ liệu Foody mà em đã thu thập được. Đồng thời, so sánh kết quả của các mô hình với nhau.

3.1. Mô hình RSA-SM

Mô hình thứ nhất là reliable sentiment analysis stack max (RSA-SM) được mô tả như hình 3-1. Mô hình được lấy ý tưởng từ bài báo của Macro Seeland [19], khoá luận lấy ý tưởng stack và max các đặc trưng của ảnh để có được một đặc trưng mới sau đó em sẽ kết hợp với đặc trưng văn bản để phù hợp yêu cầu bài toán. Các đặc trưng của tập ảnh trong một bình luận sẽ được xếp chồng và lấy giá trị lớn nhất tại mỗi điểm ảnh trong tập đặc trưng ảnh để thu được một đặc trưng ảnh đại diện cho tập ảnh đó.



Hình 3-1: Mô hình RSA-SM

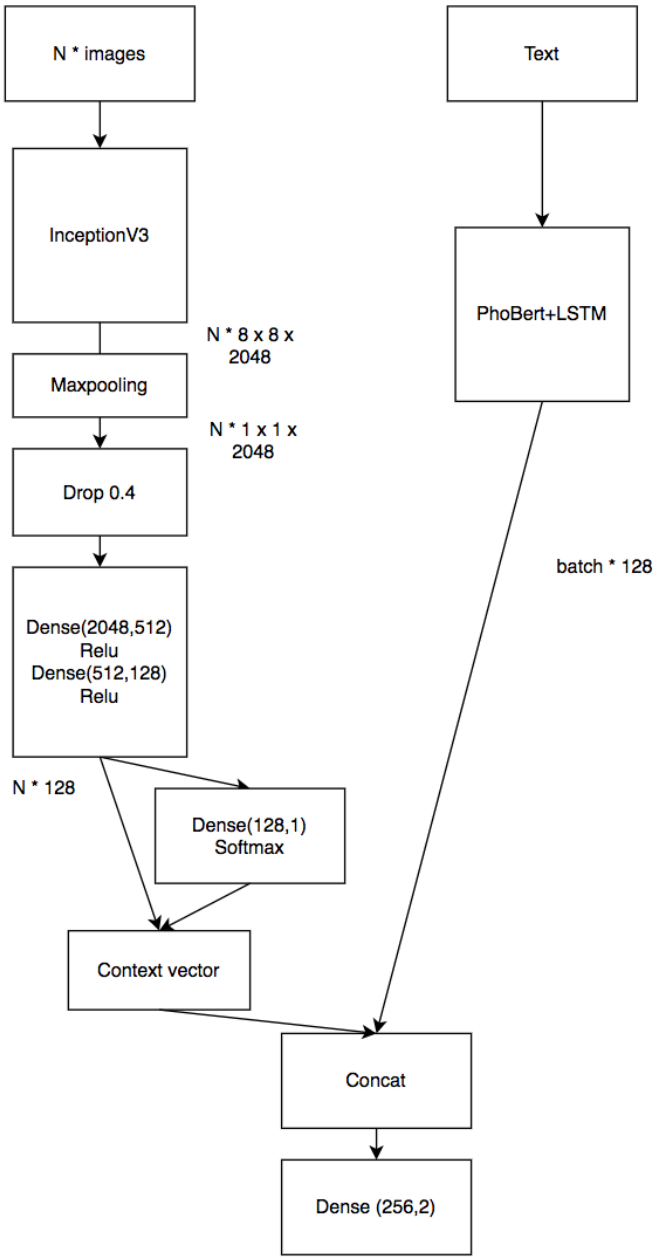
3.2. Mô hình RSA-ALF

Với mô hình reliable sentiment analysis attention late fustion (RSA-ALF), mô hình được miêu tả như hình 3-2. Ý tưởng của mô hình được lấy cảm hứng từ bài báo của Ignazio Gallo [20] sử dụng các pretrained model như Inception-V3, thay BERT bằng PhoBERT và sử dụng LSTM sau đó em cải tiến kết hợp các đặc trưng bằng cơ chế attention. Mô hình sử dụng InceptionV3 cho quá trình trích xuất đặc trưng ảnh bởi độ chính xác cao và cũng như ưu điểm trong việc tính toán của mô hình này so với các mô hình tích chập khác. InceptionV3 sẽ được loại bỏ hai tầng cuối cùng là Average pooling và lớp phân loại (fully connected) và thay vào đó em sẽ sử dụng MaxPooling 8x8 và

Drop out với xác suất là 40%. Đặc trưng ảnh sau đó sẽ được lần lượt cho qua các lớp fully connected 2048, 512, 128 kết hợp cùng hàm kích hoạt biến đổi phi tuyến Relu. Các đặc trưng ảnh sau khi được đưa về kích thước là 128 sẽ được áp dụng kĩ thuật attention bằng cách cho qua lớp fully connected và softmax để tính trọng số về độ quan trọng của các ảnh và các hệ số sẽ được tiếp tục nhân với các đặc trưng ảnh tương ứng để sinh ra vector ngữ cảnh của ảnh.

Đối với đặc trưng văn bản mô hình sử dụng $\text{PhoBERT}_{\text{base}}$ kết hợp LSTM với đơn vị ẩn (hidden unit) là 128. Câu bình luận sau khi qua PhoBERT ta thu được embedding của mỗi token trong câu và chuỗi embedding token của cả văn bản sẽ là đầu vào của mạng LSTM. Sau khi qua LSTM, embedding của token cuối sẽ được sử dụng là vector đặc trưng đại diện cho ý nghĩa cả văn bản.

Hai đặc trưng ảnh và văn bản sẽ được nối lại (concat) thành vector có số chiều là 256 và vector này sẽ được cho qua lớp classifier với số nút là 2 tương ứng với số nhãn mà mô hình phải dự đoán.

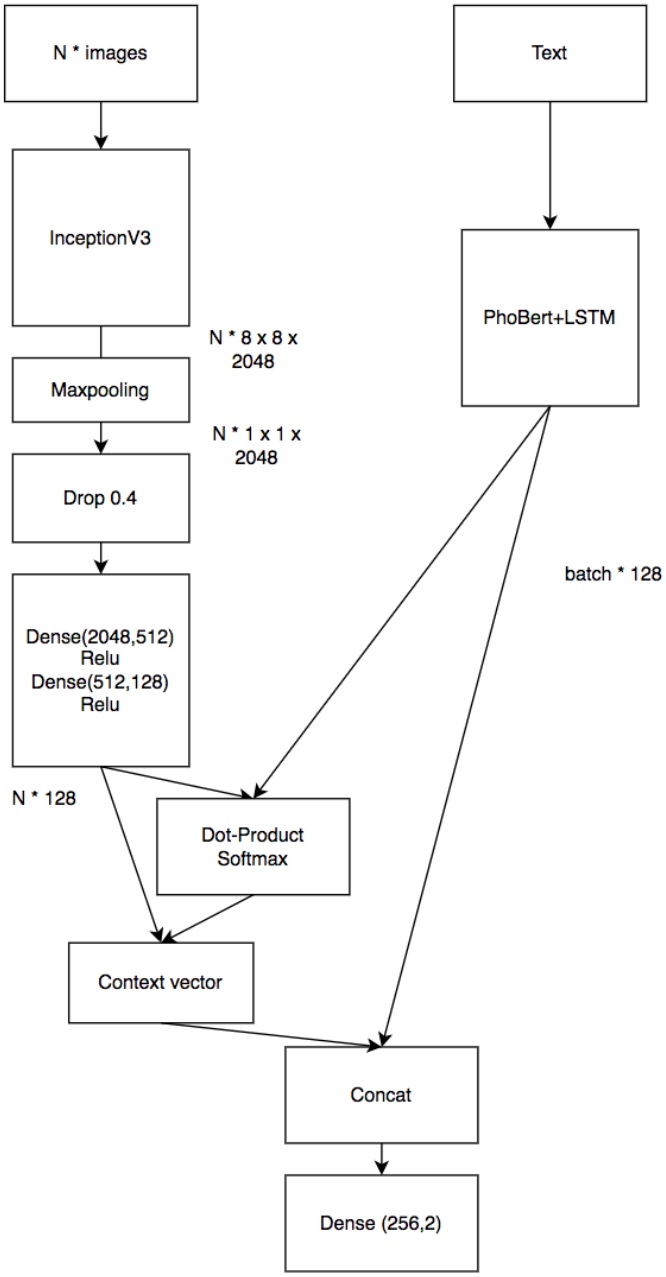


Hình 3-2: Mô hình RSA-ALF

3.3. Mô hình RSA-AEF

Kĩ thuật late fusion trong mô hình RSA-ALF chưa thực sự xử lí được tốt nhất mối quan hệ của đặc trưng ảnh và văn bản vì vậy em sẽ áp dụng kĩ thuật early fusion cho mô hình reliable sentiment analysis attention early fusion (RSA-AEF) bằng cách sử dụng cơ chế attention để tính sự tương quan giữa đặc trưng bình luận với đặc trưng của các ảnh trong bình luận đó. Các đặc trưng ảnh sẽ được nhân vô hướng với đặc trưng của văn

bản và đưa qua lớp softmax để có được các trọng số tương ứng. Các hệ số đó sẽ được nhân tương ứng với các đặc trưng ảnh để sinh ra vector ngữ cảnh của tập ảnh.



Hình 3-3: Mô hình RSA-AEF

Mô hình RSA-AEF được mô tả như hình 3-3. Cách xử lí của mô hình RSA-AEF với mô hình InceptionV3 và mô hình PhoBERT, LSTM giống với cách xử lí của mô

hình RSA-ALF chỉ khác nhau ở cách sử dụng cơ chế attention để tính trọng số cho mỗi ảnh trong bình luận.

3.4. Mô tả dữ liệu

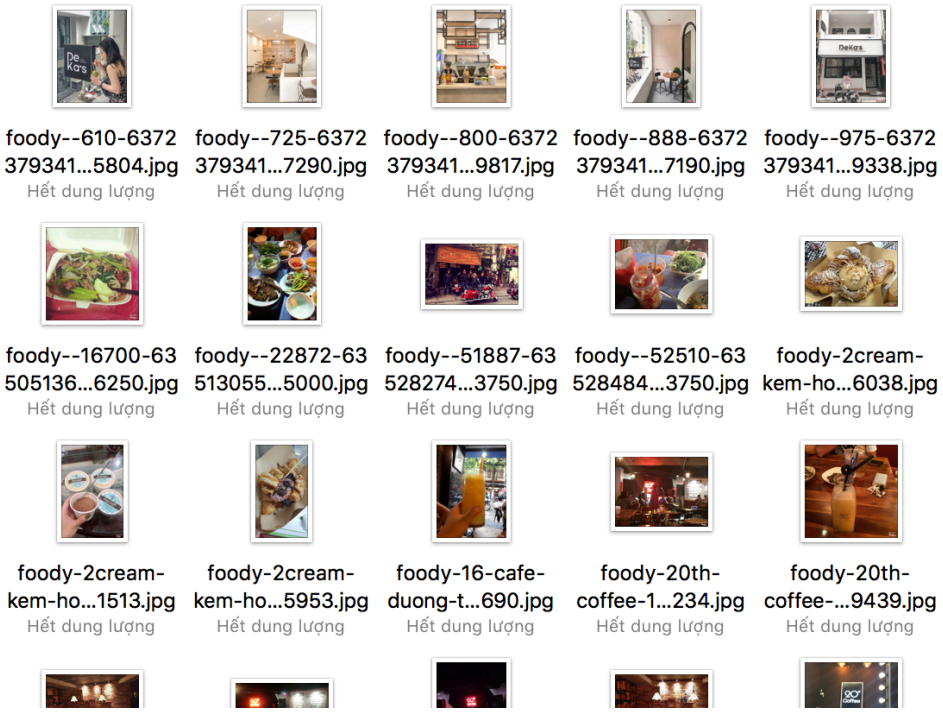
3.4.1. Tập dữ liệu Foody

Tập dữ liệu Foody mà em thu thập là tập dữ liệu các bình luận của người dùng trên trang ẩm thực https://www.foody.vn/ – cộng đồng tin cậy cho mọi người có thể tìm kiếm, đánh giá, bình luận các địa điểm ăn uống. Dữ liệu sau khi thu thập được lưu dưới dạng tệp 'csv' với hơn 9,000 bình luận. Mỗi dòng dữ liệu sẽ bao gồm bình luận (Comment), đường dẫn của ảnh (image_urls) và số điểm của mỗi bình luận (score).

	ResId	Comment	image_urls	score
0	3648046	Đặt dồi sụn của quán vì đọc comment thấy hấp d...	local1_folder-1/foody-doi-sun-pate-shop-online...	5.8
1	3695359	Dồi khá ngon, mua về còn nóng mở ra thơm phức,...	local1_folder-1/foody-doi-sun-pate-shop-online...	9.0
2	3695487	Đặt xuất mỳ trộn thập cẩm, khá đầy đủ và đầy đ...	local1_folder-1/foody-doi-sun-pate-shop-online...	9.4
3	4256913	Không hiểu sao quán này được 7.9 luôn. Đặt bán...	local1_folder-1/foody-doi-sun-pate-shop-online...	4.6
4	4246644	Dồi sụn bé tẹo, giá quá cao so với các quán kh...	local1_folder-1/foody-doi-sun-pate-shop-online...	1.0
...	...	...	...	...

Hình 3-4: Hình ảnh về dữ liệu bảng Foody đã thu thập

Dữ liệu hình ảnh thu thập từ những bình luận của người dùng với hơn 20,000 ảnh được lưu trữ trong một thư mục ảnh.



Hình 3-5: Hình ảnh về dữ liệu ảnh bình luận Foody thu thập

3.4.2. Tiền xử lý dữ liệu

Các bình luận trên dữ liệu Foody sau khi thu thập về còn là dữ liệu thô và chứa nhiều thông tin gây nhiễu cho mô hình vì vậy việc tiền xử lí và làm sạch dữ liệu là giai đoạn cần thiết. Các bước tiền xử lí dữ liệu văn bản sẽ bao gồm các bước sau:

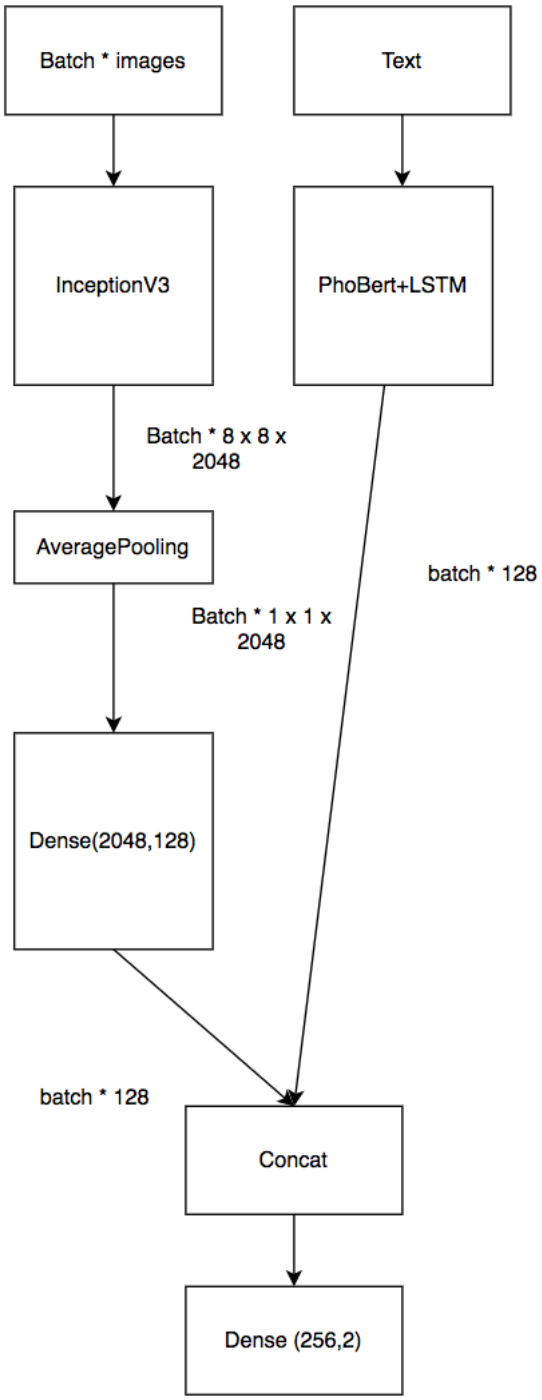
- Bước đầu em sẽ chuẩn hoá Unicode tiếng việt, do hiện nay các chuẩn đánh máy có thể khác nhau giữa các máy tính của người dùng, 2 loại mã Unicode phổ biến là tổ hợp và dựng sẵn. Nếu các kí tự không được chuẩn hoá thì cùng một từ với hai loại mã khác nhau thì khi đưa vào mô hình sẽ xem đó là hai từ khác nhau. Vì vậy em sẽ chuẩn hoá về cùng một mã Unicode dựng sẵn do mã này phổ biến hơn.
- Xoá các kí tự đặc biệt như các dấu câu, các biểu tượng cảm xúc (emoji) của người dùng sử dụng trong quá trình bình luận bằng phương pháp regular expression
- Xoá các khoảng trắng dấu cách liên tiếp nhau
- Đưa tất cả kí tự về kí tự viết thường

- Do PhoBert sử dụng rdrsegmenter từ thư viện VNcorenlp trong quá trình huấn luyện trước vì vậy em sẽ sử dụng công cụ này để tách từ.

3.5. Các mô hình so sánh

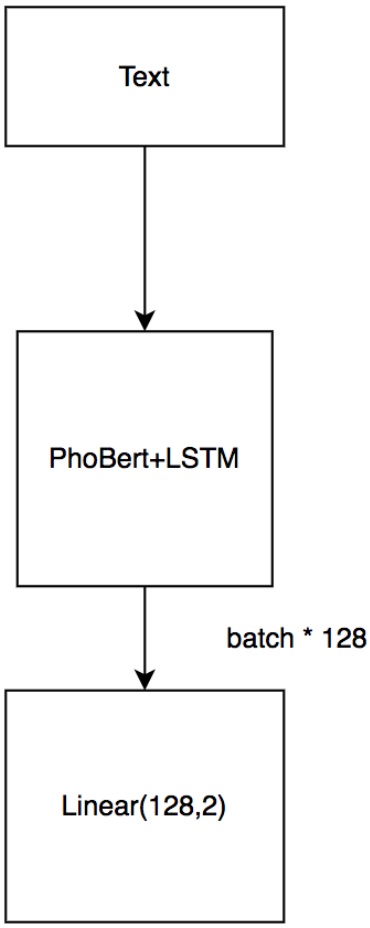
Em sẽ tiến hành đánh giá độ hiệu quả của các mô hình so với nhau. Em sẽ so sánh các mô hình đề xuất đa phương thức (multimodal) với các mô hình cơ sở là mô hình Voting-multimodal, mô hình chỉ sử dụng văn bản PhoBERT+LSTM (text-only) và mô hình chỉ sử dụng ảnh Attention-InceptionV3 (image-only)

- Voting Multimodal: Mô hình cơ sở được lấy ý tưởng từ bài báo của Ignazio Gallo [20] trong bài báo mô hình chỉ cho một ảnh và một bình luận đi qua mô hình và em sẽ sử dụng cơ chế voting để xử lí với nhiều ảnh và một bình luận. Mô hình được mô tả như hình 3-6. Tập bao gồm N ảnh và bình luận sẽ được nhân lặp lại tách thành N ảnh + N bình luận và đưa qua mô hình. Trong giai đoạn kiểm thử sử dụng cơ chế voting với nhãn đoán được là nhãn chiếm đa số trong N ảnh + N bình luận.



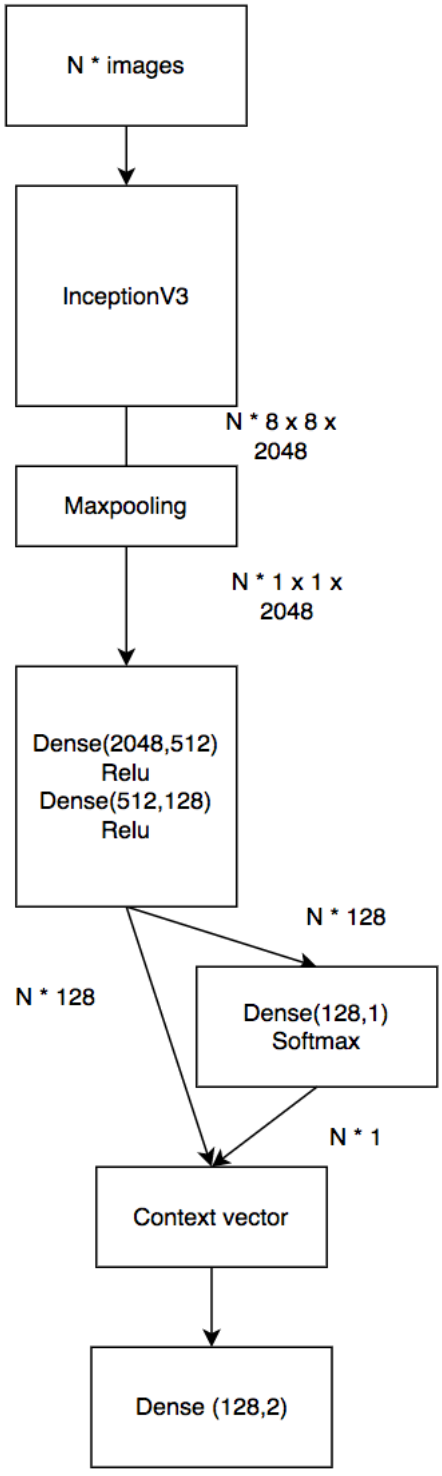
Hình 3-6: Mô hình Voting-multimodal

- PhoBERT+LSTM (text-only): mô hình được tách riêng từ mô hình đa phương thức và cài đặt các siêu tham số và huấn luyện như mô hình đa phương thức. Mô hình được mô tả như hình 3-7



Hình 3-7: Mô hình Phobert+LSTM

- Mô hình Attention-InceptionV3 (image-only): mô hình cũng được tách riêng từ mô hình đa phương thức và cài đặt các siêu tham số và huấn luyện như mô hình đa phương thức. Mô hình được mô tả như hình 3-8



Hình 3-8: Mô hình Attention-InceptionV3

3.6. Cài đặt thử nghiệm

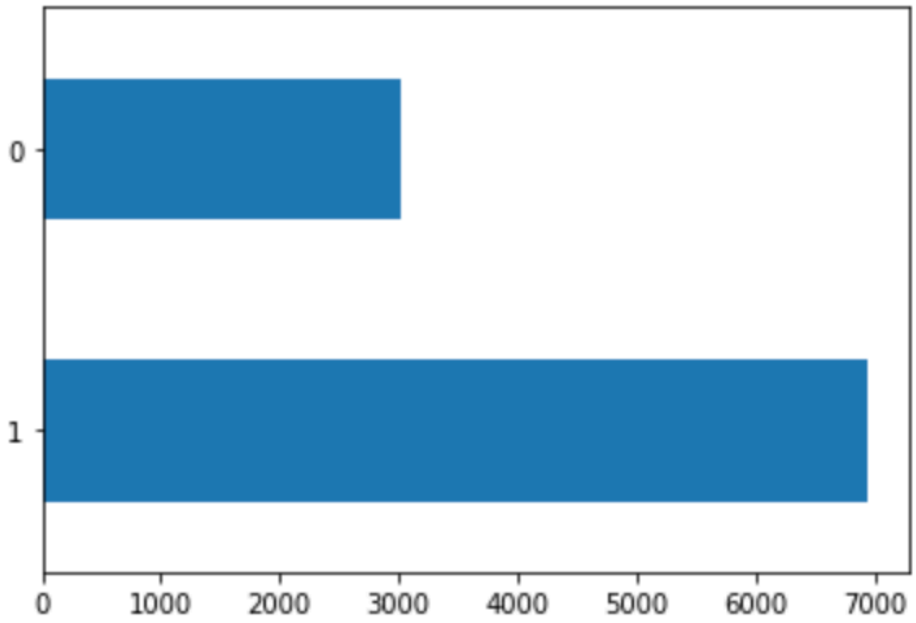
Tập dữ liệu được chia thành 3 tập là tập huấn luyện, tập đánh giá và tập kiểm thử (train –validation – test). Tập huấn luyện, tập đánh giá, tập kiểm thử được chia theo tham số 8-1-1 trên toàn bộ tập dữ liệu.

Tập đánh giá sẽ được sử dụng cho phương pháp earlystopping quá trình huấn luyện sẽ được dừng lại khi hàm mất mát trên tập dữ liệu đánh giá có xu hướng tăng hoặc có xu hướng đi ngang sau 8 epoch để tránh gây ra hiện tượng overfitting.

Tập đánh giá cũng sẽ được sử dụng để lưu lại kết quả tốt nhất của mô hình nếu độ đo macro f1 trên tập đánh giá là lớn nhất.

Em sử dụng thuật toán tối ưu Adam [21] với learning rate là 0,001. Adam optimizer là một thuật toán kết hợp kỹ thuật của RMSprop và Momentum. Thuật toán sử dụng hai vector momentum là momentum (m) sử dụng như kĩ thuật Momentum và squared momentum (v) sử dụng như trong kĩ thuật RMSprop cho các tham số của gradient. Nhờ khả năng tự học tham số của thuật toán Adam, nó không cần thiết kết hợp thêm một phương thức điều chỉnh learning rate để tăng tốc độ hội tụ.

Hàm mất mát sử dụng là là class weight cross entropy loss do dữ liệu khá mất cân bằng. Hình 3-9 thể hiện phân bố lệch của hai nhãn với nhãn 0 (negative), nhãn 1 (positive)



Hình 3-9: Thống kê phân bố dữ liệu

Mô hình được huấn luyện trên dịch vụ Google Colab có cấu hình như sau:

- GPU: NVIDIA Tesla K80 11GB
- RAM: 12 GB

Trong khóa luận, thử nghiệm sẽ được xây dựng trên ngôn ngữ python (python 3.10) và sử dụng framework Pytorch[2], cùng các thư viện đi kèm được giới thiệu dưới đây.

Tên thư viện	Phiên bản
numpy	1.22.3
pytorch	1.11.0
matplotlib	3.5.1
vncorenlp	1.0.3
torchvision	0.12.0

Bảng 2: Bảng các thư viện sử dụng

[2] Facebook, “Pytorch, “Facebook, [Online]. Available: https://pytorch.org/

Mô hình early fusion, late fusion và các mô hình unimodal chạy thử nghiệm trên các bộ siêu tham số được mô tả như trong bảng sau:

Siêu tham số	Mô tả	Một số lựa chọn
Batch size	Kích thước lô huấn luyện	8
Learning rate	Tốc độ học	{0.000001,0.00001, 0.0001, 0.001}
Embedding size	Số chiều của véc tơ biểu diễn một token	768
LSTM hidden units	Số chiều của véc tơ trạng thái ẩn trong LSTM	128
LSTM layers	Số lớp LSTM	1

Bảng 3: Bảng siêu tham số sử dụng trong quá trình huấn luyện

3.7. Phương pháp đánh giá

Có rất nhiều phương pháp được sử dụng để đánh giá hiệu năng cho bài toán phân tích sắc thái bình luận. Mỗi một thang đánh giá đều có ưu và nhược điểm riêng và mục tiêu mong muốn. Trong khóa luận sẽ sử dụng thang đo chính là accuracy và Macro-F1 để đánh giá hiệu năng cho việc phân loại sắc thái bình luận.

3.7.1. Thang đo accuracy

Độ chính xác là độ đo cho biết số điểm mô hình đoán đúng trên tổng số tất cả các điểm trong tập dữ liệu, được tính bằng tỉ lệ giữa số điểm dự đoán chính xác và tổng số điểm trong tập dữ liệu kiểm thử. Công thức được tính như sau:

$$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

Trong đó True Positive là số lượng các phần tử được đoán đúng lớp positive, True Negative là số lượng các phần tử được đoán đúng lớp negative, False Positive là số lượng các phần tử được đoán sai lớp positive, False Negative là số lượng các phần tử được đoán sai lớp negative

3.7.2. Thang đo Macro-F1

Thang đo accuracy là một thang đo đơn giản giúp đánh giá được hiệu suất của mô hình tuy nhiên thang đo này sẽ bị tác động do dữ liệu bị mất cân bằng bởi mô hình sẽ có thiên hướng dự đoán thiên lệch về nhãn đa số vì vậy mặc dù mô hình không tốt nhưng thang đo accuracy lại đạt giá trị cao dẫn tới đánh giá sai chất lượng mô hình. Thang đo macro-f1 giải quyết vấn đề này của accuracy. Thang đo macro-f1 được kết hợp từ hai thang đo khác là precision và recall. Thang đo recall là tỉ lệ mô hình đoán đúng trên tổng số nhãn đúng thực sự, còn precision là tỉ lệ mô hình đoán đúng trên tổng số dự đoán mô hình đưa ra. Các giá trị Recall, Precision, Macro-F1 được tính theo công thức sau:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\frac{2}{\text{Macro} - \text{F1}} = \frac{1}{\text{Macro} - \text{precision}} + \frac{1}{\text{Macro} - \text{recall}}$$

Với Macro-precision và macro-recall là trung bình cộng của precision và recall theo các lớp.

3.8. Thực nghiệm và đánh giá kết quả

3.8.1. So sánh các mô hình đề xuất

Các mô hình sẽ được chạy thực nghiệm và đánh giá trên 5 seed value.

Theo thang đo accuracy ta có kết quả thực nghiệm theo bảng sau:

Model / **Seed**	RSA-SM	RSA-ALF	RSA-AEF
0	0,9438	0,9418	0,9538
7	0,9508	0,9317	0,9438
66	0,9438	0,9538	0,9458

25	0,9358	0,9488	0,9438
84	0,9428	0,9368	0,9468
Trung bình	0,9434	0,9426	**0,9468**

Bảng 4: Bảng so sánh accuracy giữa các mô hình đề xuất

Nhận xét: Sau 5 seed thử nghiệm ta thu được kết quả trung bình với kết quả cao nhất đạt được từ mô hình RSA-AEF (94,68%) theo sau lần lượt là RSA-SM, RSA-ALF.

Theo thang đo Macro-f1 ta có kết quả thực nghiệm theo bảng sau:

Model / **Seed**	RSA-SM	RSA-ALF	RSA-AEF
0	0,9327	0,9312	0,9441
7	0,9410	0,9196	0,9323
66	0,9331	0,9448	0,9352
25	0,9241	0,9389	0,9329
84	0,9319	0,9247	0,9362
Trung bình	0,9325	0,9318	**0,9361**

Bảng 5: Bảng so sánh macro-f1 giữa các mô hình đề xuất

Nhận xét: Sau 5 seed thử nghiệm ta thu được kết quả trung bình với kết quả cao nhất đạt được từ mô hình RSA-AEF (93,61%) theo sau lần lượt là RSA-SM, RSA-ALF.

Có thể thấy được việc sử dụng cơ chế attention giữa ảnh và text theo kĩ thuật kết hợp trước (earlyfusion) trong mô hình RSA-AEF đã mang lại kết quả tốt hơn so với việc chỉ dùng cơ chế attention cho riêng ảnh trong mô hình RSA-ALF rồi nối hai đặc trưng một cách độc lập (latefusion). Mô hình RSA-AEF cũng có kết quả tốt hơn so với RSA-

SM - một mô hình cũng chỉ xử lí đặc trưng ảnh riêng và kết hợp sau (latefusion). Điều này chứng tỏ rằng phương pháp early fusion đã mang lại sự kết hợp giữa hai đặc trưng ảnh và văn bản một cách chặt chẽ hơn thay vì chỉ xử lí các đặc trưng một cách độc lập rồi nối lại với nhau của phương pháp late fusion.

3.8.2. So sánh các mô hình cơ sở

Các mô hình sẽ được chạy thực nghiệm và đánh giá trên 5 seed value.

Theo thang đo accuracy ta có kết quả thực nghiệm theo bảng sau:

Model / **Seed**	PhoBERT+LSTM (text-only)	Attention-InceptionV3 (image-only)	Voting-Multimodal	RSA-AEF
0	0,9398	0,7261	0,9307	0,9538
7	0,9448	0,7241	0,9468	0,9438
66	0,9498	0,6750	0,9438	0,9458
25	0,9398	0,7211	0,9428	0,9438
84	0,9368	0,7191	0,9388	0,9468
Trung bình	0,9422	0,7131	0,9405	**0,9468**

Bảng 6: Bảng so sánh accuracy giữa các mô hình cơ sở

Nhận xét: Sau 5 seed thử nghiệm ta thu được kết quả trung bình với kết quả cao nhất đạt được từ mô hình RSA-AEF (94,68%) theo sau lần lượt là Phobert+LSTM, Voting-Multimodal, Attention-InceptionV3.

Theo thang đo Macro-f1 ta có kết quả thực nghiệm theo bảng sau:

Model / **Seed**	PhoBERT+LSTM (text-only)	Attention-InceptionV3 (image-only)	Voting-Multimodal	RSA-AEF
0	0,9290	0,6778	0,9194	0,9441
7	0,9336	0,6864	0,9367	0,9323
66	0,9400	0,6447	0,9325	0,9352
25	0,9287	0,6746	0,9319	0,9329
84	0,9258	0,6783	0,9278	0,9362
Trung bình	0,9314	0,6724	0,9296	**0,9361**

Bảng 7:Bảng so sánh macro-f1 giữa các mô hình cơ sở

Nhận xét: Sau 5 seed thử nghiệm ta thu được kết quả trung bình với kết quả cao nhất đạt được từ mô hình RSA-AEF (93,61%) theo sau lần lượt là Phobert+LSTM, Voting-Multimodal, Attention-InceptionV3.

Có thể thấy được việc kết hợp đặc trưng ảnh và văn bản trong mô hình đa phương thức đã mang lại sự tin cậy và chính xác hơn các mô hình đơn phương thức trong tác vụ phân loại này. Việc sử dụng cơ chế attention giữa ảnh và text (earlyfusion) trong mô hình RSA-AEF đã mang lại kết quả tốt hơn so với so với mô hình cơ sở VotingMultimoddal và các mô hình đơn phương thức.

3.8.3. Một số thực nghiệm định tính

Trong phần này, em sẽ phân tích một số kết quả định tính giúp chúng ta thấy được việc kết hợp đặc trưng của ảnh và văn bản của mô hình đa phương thức cụ thể là mô hình RSA-AEF giúp nâng cao tính tin cậy trong việc phân loại tốt những trường hợp gây nhiễu cho mô hình đơn phương thức.

- Một số kết quả chống nhiễu mô hình RSA-AEF so với mô hình Phobert+LSTM

<table>
<tr><td colspan="2">Ảnh người dùng chụp:
  </td></tr>
<tr><td colspan="2">Văn bản: Ngon lắm nha, mình thấy nhiều bạn bị đổ nhân ra ngoài là do shipper cả thôi</td></tr>
<tr><td colspan="2">Nhãn: Positive</td></tr>
<tr><td>PhoBERT+LSTM dự đoán: Negative</td><td>RSA-AEF dự đoán: Positive</td></tr>
</table>

Nhận xét: Đối với ví dụ trên câu bình luận với một số từ mang nghĩa tiêu cực gây cho mô hình PhoBERT+LSTM bị phân loại nhầm sang nhãn Negative tuy nhiên mô hình RSA-AEF lại dự đoán chính xác nhãn là Positive dựa vào sự kết hợp ảnh và văn bản với hình ảnh có thể thấy được món ăn mang nhiều màu sắc và nhìn ngon miệng.

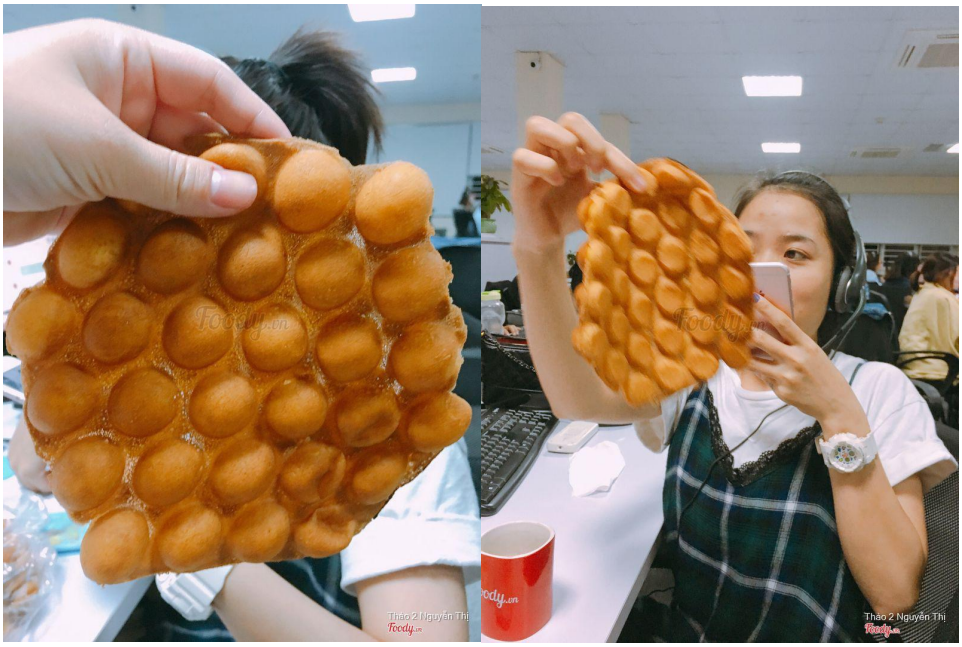
<table>
<tr><td colspan="2">Ảnh người dùng chụp:
  </td></tr>
<tr><td colspan="2">Văn bản: cả quán 5 nhân_viên lác_đác vài khách mà đợi hơn 20 mới được 1 cốc cafe phân_nửa các bạn nhân_viên ưu_ái lấy hẳn ống hút to cho đồ nóng tí bỏng gọi bạc sỉu nóng thì lấy nhầm thành nước socola pha loãng mình gọi đúng thế vì ngoại_trừ màu nâu thì là nước_lọc có mùi socola chê xong đến khen nhé quán có món cafe chuối khá lạ hương_vị chuối khá thơm và cafe khá đậm sánh dù hơi ngọt giá_như các bạn phục_vụ nhanh hơn cốc đầy_đặn hơn giá tăng thêm chút cũng đc thì mình nhất_định sẽ trung_thành vs quán</td></tr>
<tr><td colspan="2">Nhãn: Negative</td></tr>
<tr><td>PhoBERT+LSTM dự đoán: Positive</td><td>RSA-AEF dự đoán: Negative</td></tr>
</table>

Nhận xét: Đối với ví dụ trên câu bình luận gây cho mô hình PhoBERT+LSTM bị phân loại nhầm sang nhãn Positive tuy nhiên mô hình RSA-AEF lại dự đoán chính xác nhãn là Negative dựa vào sự kết hợp ảnh và văn bản với hình ảnh có thể thấy được chụp không được chỉnh chu thiếu màu sắc.

- Một số kết quả chống nhiễu mô hình RSA-AEF so với mô hình Attention-InceptionV3

<table>
<tr><td colspan="2">Ảnh người dùng chụp:
  </td></tr>
<tr><td colspan="2">Văn bản: mình mua cả 3 loại bánh gà thường phô_mai và cari bánh gà đầy_đặn vỏ bánh không bị dày như 1 số quán khác cả 3 loại đều ngon</td></tr>
<tr><td colspan="2">Nhãn: Positive</td></tr>
<tr><td>Attention-Inceptionv3 dự đoán:
Negative</td><td>RSA-AEF dự đoán: Positive</td></tr>
</table>

Nhận xét: Đối với ví dụ trên ảnh đầu vào gây cho mô hình Attention-InceptionV3 bị phân loại nhầm sang nhãn Negative, có thể thấy những hình ảnh chụp trong điều kiện thiếu sáng không chỉnh chu làm ảnh hưởng đến kết quả phân loại tuy nhiên mô hình RSA-AEF lại dự đoán chính xác nhãn là Positive dựa vào sự kết hợp ảnh và văn bản.

<table>
<tr><td colspan="2">Ảnh người dùng chụp:
  </td></tr>
<tr><td colspan="2">Văn bản: em ăn bánh trứng gà non khá nhiều nơi nhưng nói thật em không hài_lòng cho lắm với bánh trứng gà non ở quán này cho lắm hôm em đặt về ăn là lúc em như một con ma đói mà những lúc đói thì người_ta hay bảo dù đồ có chán thì ăn cũng thấy ngon nhưng thề là em ăn thấy chẳng ngon_miệng gì cả pải nói thế_nào nhờ bột dầy không được xốp không thơm ngậy và bánh của em nướng còn bị cháy cơ em là đứa hơi kĩ_tính nên chắc là sẽ không có lần thứ hai em ăn_ở quán này nữa</td></tr>
<tr><td colspan="2">Nhãn: Negative</td></tr>
<tr><td>Attention-Inceptionv3 dự đoán:
Positive</td><td>RSA-AEF dự đoán: Negative</td></tr>
</table>

Nhận xét: Đối với ví dụ trên ảnh đầu vào gây cho mô hình Attention-InceptionV3 bị phân loại nhầm sang nhãn Positive tuy nhiên mô hình RSA-AEF lại dự đoán chính xác nhãn là Negative.

Chương 4.Tổng kết

4.1. Kết luận

Khoá luận đưa ra những nghiên cứu về dữ liệu bình luận của người dùng trên trang ẩm thực Foody. Từ đó đề xuất 3 mô hình đa phương thức là RSA-SM, RSA-ALF, RSA-AEF nhằm nâng cao tính tin cậy và chính xác cho tác vụ phân loại sắc thái bình luận với ý tưởng chính mang lại kết quả tốt đó là sử dụng cơ chế attention và các phương pháp early fusion trong mô hình RSA-AEF để kết hợp hai đặc trưng giữa ảnh và văn bản. Các mô hình được huấn luyện với hơn 9,000 bình luận và 20,000 ảnh từ dữ liệu trang Foody và mô hình RSA-AEF đạt kết quả tốt nhất với 94,68% trên thang đo accuracy và 93,61% trên thang đo Macro-f1. Mô hình RSA-AEF cũng được so sánh với mô hình đa phương thức cơ sở và đơn phương thức cho ảnh cũng như văn bản và thể hiện được tính tin cậy và chính xác so với các mô hình trong tác vụ phân tích sắc thái bình luận. Mặc dù vậy khoá luận vẫn còn những hạn chế: chưa xây dựng được một hệ thống hoàn chỉnh, kết quả phân loại vẫn chưa thực sự tốt.

4.2. Hướng phát triển trong tương lai

Mô hình hiện tại vẫn chưa có kết quả quá ấn tượng trong tác vụ phân loại sắc thái bình luận này. Trong tương lai em dự định thử nghiệm nhiều phương pháp hơn nữa để kết hợp các đặc trưng trong mô hình đa phương thức tốt hơn và thử nghiệm trên bộ dữ liệu lớn hơn. Và hướng tới sẽ xây dựng một hệ thống hoàn chỉnh, tích hợp mô hình lên các nền tảng thương mại điện tử.

Chương 5. Tài liệu tham khảo

[1] Saria, Suchi, and Adarsh Subbaswamy. “Tutorial: safe and reliable machine learning.” arXiv preprint arXiv:1904.07204(2019).

[2] Nasukawa, Tetsuya, and Jeonghee Yi. "Sentiment analysis: Capturing favorability using natural language processing." Proceedings of the 2nd international conference on Knowledge capture. 2003.

[3] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." arXiv preprint cs/0205070 (2002).

[4] Kieu, Binh Thanh, and Son Bao Pham. "Sentiment analysis for Vietnamese." 2010 Second International Conference on Knowledge and Systems Engineering. IEEE, 2010.

[5] Le, Lac Si, et al. "A Multi-filter BiLSTM-CNN Architecture for Vietnamese Sentiment Analysis." International Conference on Computational Collective Intelligence. Springer, Cham, 2020.

[6] https://analyticsmitra.wordpress.com/2018/02/05/artificial-neural-network-the-basic-idea-behind-machines-brain/

[7] Afshine Amidi ,Shervine Amidi, "Recurrent Neural Networks cheatsheet," [Online]. Available: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks.

[8] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[9] C. Blog, "Understanding LSTM Networks," 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[10] Nguyen, Dat Quoc, and Anh Tuan Nguyen. "PhoBERT: Pre-trained language models for Vietnamese." arXiv preprint arXiv:2003.00744 (2020).

[11] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[12] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[13] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[14] A. N. T, "Inception V3 Model Architecture," [Online]. Available: https://iq.opengenus.org/inception-v3-model-architecture/.

[15] Ian Goodfellow ,Yoshua Bengio , Aaron Courville, Deep Learning, MIT Press, 2016.

[16] Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio, "NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE," ICLR, 2015.

[17] Minh-Thang Luong, Hieu Pham,Christopher D. Manning, Sequence to sequence learning with neural networks neural machine translation, Computing Research Repository, 2015.

[18] Seeland, Marco, and Patrick Mäder. "Multi-view classification with convolutional neural networks." Plos one 16.1 (2021): e0245230.

[19] Gallo, Ignazio, et al. "Image and Text fusion for UPMC Food-101 using BERT and CNNs." 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, 2020.

[20] DP Kingma, Diederik P., and Jimmy Ba, ""Adam: A method for stochastic optimization."," arXiv preprint arXiv, 2014.