TRƯỜNG ĐẠI HỌC CÔNG NGHỆ
ĐẠI HỌC QUỐC GIA HÀ NỘI

ĐOÀN XUÂN DŨNG

TÓM TẮT VĂN BẢN SỬ DỤNG CÁC KỸ THUẬT TRONG DEEP LEARNING

Ngành: Công Nghệ Thông Tin
Chuyên ngành: Khoa học máy tính
Mã số chuyên ngành: 8480101.01

LUẬN VĂN THẠC SỸ NGÀNH CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS Nguyễn Xuân Hoài

HÀ NỘI – 2018

Lời cảm ơn

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và biết ơn đến PGS.TS Nguyễn Xuân Hoài, người thầy đã chỉ bảo và hướng dẫn tận tình trong quá trình tôi nghiên cứu khoa học và làm luận văn này.

Tôi xin chân thành cảm ơn sự giúp đỡ nhiệt tình của PGS.TS Nguyễn Lê Minh trong quá trình nghiên cứu tại Viện Khoa học và Công nghệ tiên tiến Nhật Bản (JAIST) từ tháng 4/2017 đến tháng 6/2017.

Và cuối cùng tôi xin gửi lời cảm ơn tới gia đình, người thân, bạn bè – những người luôn ở bên tôi những lúc khó khăn nhất, luôn động viên và khuyến khích tôi trong cuộc sống và trong công việc.

Tôi xin chân thành cảm ơn!

Hà Nội, ngày.....tháng.....năm 2018

Người cam đoan

Đoàn Xuân Dũng

Lời cam đoan

Tôi xin cam đoan luận văn được hoàn thành trên cơ sở nghiên cứu, tổng hợp và phát triển các nghiên cứu tóm tắt văn bản. Trong quá trình làm luận văn tôi có tham khảo các tài liệu có liên quan và đã ghi rõ nguồn gốc tài liệu.

Tôi xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Hà Nội, ngày.....tháng.....năm 2018

Người cam đoan

Đoàn Xuân Dũng

MỤC LỤC

BẢNG CÁC TỪ VIẾT TẮT

Viết tắt	**Đầy đủ**	**Ý nghĩa**
ANN	Artificial Neural Network	Mạng nơ-ron nhân tạo
FNN	Feedforward Neural Network	Mạng nơ-ron lan truyền tiến
MLP	Multilayer Perceptrons	Mạng nơ-ron đa lớp
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
LSTM	Long Short Term Memory	Mạng nơ-ron bộ nhớ ngắn dài hạn
GRU	Gated Recurrent Units	Mạng nơ-ron với các đơn vị cổng hồi quy
CNN	Convolution Neural Network	Mạng nơ-ron tích chập
BiRNN	Bi-directional Recurrent Neural Network	Mạng hai chiều RNN
Encoder-Decoder	Encoder-Decoder	Mã hóa – Giải mã

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

Mở đầu

Ngày nay, con người đang bước vào kỷ nguyên của cách mạng công nghiệp 4.0, chúng ta phải đối mặt với lượng thông tin khổng lồ trên mạng Internet. Do đó nhu cầu tóm tắt thông tin đối với mỗi văn bản là vô cùng cấp thiết. Tóm tắt văn bản là phương pháp rút gọn lại một lượng lớn các thông tin thành một bản tóm tắt ngắn gọn bởi sự lựa chọn những thông tin quan trọng và bỏ qua các thông tin dư thừa.

Thông thường tóm tắt văn bản có thể chia thành tóm tắt trích chọn (extractive summarization) và tóm tắt tóm lược (abstractive summarization). Tóm tắt trích chọn đưa ra sự tóm tắt bằng việc chọn một tập các câu trong văn bản ban đầu. Ngược lại, tóm tắt tóm lược đưa ra thông tin được thể hiện lại theo một cách khác. Tóm tắt trích chọn bao gồm các câu lấy ra từ văn bản, trong khi đó tóm tắt tóm lược sử dụng những từ và cụm từ không xuất hiện trong văn bản gốc. Tóm tắt trích chọn là phương pháp đơn giản nhưng mạnh mẽ cho tóm tắt văn bản, nó liên quan đến việc ấn định điểm số cho thành phần văn bản rồi chọn ra phần có điểm cao nhất. Tóm tắt tóm lược cần phải đọc và hiểu được văn bản để nhận thức được nội dung, sau đó tóm tắt văn bản cho ngắn gọn. Vì thế tóm tắt tóm lược cần một kỹ thuật sâu về xử lý ngôn ngữ.

Những năm gần đây chứng tỏ sự trở lại mạnh mẽ của mạng nơ-ron nhân tạo trong các mô hình học tự động với tên gọi học sâu (Deep Learning). Học sâu đã và đang được áp dụng trong nhiều bài toán khác nhau để thu được những kết quả tốt trong nhiều lĩnh vực của khoa học máy tính.

Những nghiên cứu đầu tiên cho bài toán tóm tắt văn bản sử dụng học sâu được đưa ra bởi nhóm tác giả Alexander Rush[2]. Nhóm tác giả đề xuất mô hình mạng nơ-ron attention kết hợp mô hình xác suất với một thuật toán sinh để đưa ra độ chính xác cho bài toán tóm tắt. Họ sử dụng một lượng lớn dữ liệu huấn luyện là các cặp văn bản tóm tắt, tận dụng sức mạnh của phần cứng máy tính để học ra mô hình huấn luyện. Sau đó một năm, nhóm tác giả Submit Chopra[3] mở rộng bài toán tóm tắt tới kiến trúc mạng nơ-ron hồi quy – RNN. Kết quả đạt tốt nhất trên tập Gigaword và DUC-2004. Tiếp đó, nhóm của Ramesh Nallapti [19] đưa ra bản tóm tắt sử dụng mạng RNN Attention Encoder-Decoder. Kết quả đạt cao nhất trên hai bộ dữ liệu khác nhau.

Gần đây, tác giả Nguyễn Viết Hạnh [25] đã nghiên cứu vấn đề tóm tắt văn bản sử dụng mô hình LSTM trong học sâu, áp dụng cho cả tiếng Anh và tiếng Việt. Kết quả tác giả đưa ra cho thấy hiệu quả của các mô hình học sâu đối với bài toán này.

Mạng nơ-ron tích chập (CNN) đã được áp dụng thành công trong các lĩnh vực của xử lý ảnh, xử lý video. Trong xử lý ngôn ngữ tự nhiên, Yoo Kim[5] đã áp dụng nâng cao kết quả bài toán phân tích cảm xúc và phân loại câu hỏi. Nhóm Nal Kalchbrenner[6] mô tả kiến trúc CNN động cho bài toán gán nhãn ngữ nghĩa câu. Yoo Kim[7] đưa ra một kiến trúc mô hình nơ-ron đơn giản kết hợp mạng nơ-ron tích chập và mạng highway trên ký tự của câu. Tiếp theo đó, nhóm tác giả Jason Lee[8] giới thiệu mạng ký tự convolution với max pooling để mã hóa giảm chiều dài của câu trình bày. Kết quả của họ chứng tỏ mô hình ký tự cho kết quả cao hơn các mô hình trong dịch máy hiện tại.

Với những thành công của mạng nơ-ron tích chập trong xử lý ngôn ngữ tự nhiên, tôi muốn cài đặt mạng nơ-ron tích chập và các mô hình trong Deep learning vào bài toán tóm tắt văn bản, kết quả trên tập dữ liệu Gigaword và DUC cho thấy hiệu quả của phương pháp này.

Ngoài phần mở đầu và phần kết luận, luận văn được chia thành 4 chương như sau:

Chương 1: Giới thiệu bài toán tóm tắt văn bản. Trình bày khái niệm và các phương pháp tiếp cận cho bài toán.

Chương 2: Cơ sở lý thuyết. Trình bày những khái niệm và mô hình trong học sâu.

Chương 3: Mô hình đề xuất. Trình bày cơ chế attention cùng thuật toán tìm kiếm chùm và áp dụng vào mô hình đề xuất.

Chương 4: Thực nghiệm và đánh giá. Trình bày quá trình thử nghiệm và đưa ra một số đánh giá, nhận xét cùng kết quả đạt được.

Chương 1: Giới thiệu tóm tắt văn bản

Tóm tắt văn bản là quá trình trích rút những thông tin quan trọng nhất từ một văn bản để tạo ra phiên bản ngắn gọn, xúc tích mang đầy đủ lượng thông tin của văn bản gốc kèm theo đó là tính đúng đắn về ngữ pháp và chính tả. Bản tóm tắt phải giữ được những thông tin quan trọng của toàn bộ văn bản chính. Bên cạnh đó, bản tóm tắt cần phải có bố cục chặt chẽ có tính đến các thông số như độ dài câu, phong cách viết và cú pháp văn bản.

Phụ thuộc vào số lượng các văn bản, kỹ thuật tóm tắt có thể chia làm hai lớp: đơn văn bản và đa văn bản. Tóm tắt đơn văn bản chỉ đơn giản là rút gọn một văn bản thành một sự trình bày ngắn gọn. Trong khi đó tóm tắt đa văn bản phải rút gọn một tập các văn bản thành một sự tóm tắt. Tóm tắt đa văn bản có thể xem như một sự mở rộng của tóm tắt đơn văn bản và thường dùng với thông tin chứa trong các cụm văn bản, để người dùng có thể hiểu được cụm văn bản đó. Tóm tắt đa văn bản phức tạp hơn tóm tắt đơn văn bản vì phải làm việc trên số lượng văn bản nhiều hơn.

Xét về phương pháp thực hiện, tóm tắt văn bản có hai hướng tiếp cận là tóm tắt theo kiểu trích chọn – "extraction" và tóm tắt theo kiểu tóm lược ý – "abstraction". Phương pháp tóm tắt trích chọn là công việc chọn ra một tập con những từ đã có, những lời nói hoặc những câu của văn bản gốc để đưa vào khuôn mẫu tóm tắt. Ngược lại phương pháp tóm tắt tóm lược xây dựng một biểu diễn ngữ nghĩa bên trong và sau đó sử dụng kỹ thuật xử lý ngôn ngữ để tạo ra bản tóm tắt gần gũi hơn so với những gì con người có thể tạo ra. Bản tóm tắt như vậy có thể chứa những từ không có trong bản gốc. Nghiên cứu về phương pháp tóm tắt tóm lược là một bước tiến quan trọng và tạo sự chủ động, tuy nhiên do các ràng buộc phức tạp nên các nghiên cứu cho đến nay chủ yếu tập trung vào phương pháp tóm tắt trích chọn. Trong một vài lĩnh vực ứng dụng, phương pháp tóm tắt trích chọn đem lại nhiều tri thức hơn.

Một lượng lớn các cách tiếp cận để xác định nội dung quan trọng cho việc tự động tóm tắt được phát triển tới ngày nay. Cách tiếp cận chủ đề đầu tiên nhận một biểu diễn trung gian của văn bản để đạt được chủ đề thảo luận. Dựa vào những sự biểu diễn này, các câu trong văn bản đầu vào được ghi điểm theo độ quan trọng. Theo một cách tiếp cận khác, văn bản được biểu diễn bởi một tập các thuộc tính

cho độ quan trọng mà không nhằm xác định chủ đề. Các thuộc tính thông thường được kết nối lại sử dụng các kỹ thuật học máy, giúp việc xác định điểm số cho độ quan trọng trong câu. Cuối cùng, một bản tóm tắt được sinh ra bằng việc lựa chọn các câu theo một cách tham lam. Việc chọn các câu được thực hiện trong một tóm tắt 1-1 hoặc bằng lựa chọn tối ưu toàn cục để chọn ra tập các câu tốt nhất cho bản tóm tắt. Sau đây xin đưa ra một cách nhìn tổng quan trên các khía cạnh với các cách biểu diễn, cách tính điểm hoặc lựa chọn chiến lược tóm tắt đảm bảo hiệu quả của bản tóm tắt.

1.1. Tóm tắt trích chọn [1]

Hệ thống tóm tắt cần đưa ra bản tóm tắt ngắn gọn và trôi chảy chứa đựng những thông tin thiết yếu của văn bản đầu vào. Trong phần này tôi thảo luận về các hệ thống tóm tắt trích chọn để đưa ra các đoạn văn ngắn và giải thích hiệu quả tóm tắt. Những bản tóm tắt xác định các câu quan trọng trong đầu vào, có thể là một văn bản hoặc một tập các văn bản liên quan và kết nối chúng với nhau thành một bản tóm tắt. Sự quyết định xung quanh nội dung nào là quan trọng trước hết hướng về đầu vào của bản tóm tắt.

Sự lựa chọn tập trung vào tóm tắt trích chọn bỏ qua một lượng lớn văn bản sinh ra bởi tóm tắt tóm lược, nhưng cho phép chúng ta tập trung vào các cách tiếp cận vượt trội để dễ dàng điều chỉnh thông tin người dùng quan tâm cho đơn văn bản và đa văn bản. Hơn nữa, bằng kiểm tra các giai đoạn trong sự hoạt động của bản tóm tắt, chúng ta có thể tập trung vào sự tương đồng và sự khác biệt trong các cách tiếp cận tóm tắt, liên quan tới các thành phần cốt yếu của hệ thống và có thể giải thích cho điểm ưu việc của kỹ thuật lựa chọn so với các kỹ thuật khác.

Để hiểu hơn về sự điều khiển các hệ thống tóm tắt và để nhấn mạnh các lựa chọn hệ thống thiết kế cần làm, tôi phân biệt ba nhiệm vụ độc lập tương đối thực hiện bởi tất cả các bản tóm tắt: Khởi tạo sự biểu diễn trung gian cho đầu vào để đạt được các khía cạnh quan trọng nhất của văn bản, ghi điểm cho câu dựa vào sự trình diễn và lựa chọn một bản tóm tắt chứa các câu văn.

1.1.1. Giai đoạn trình diễn trung gian

Cách tiếp cận biểu diễn chủ đề chuyển đổi văn bản tới một sự biểu diễn trung gian hiểu như chủ đề của văn bản. Các phương pháp tóm tắt phổ biến nhất dựa vào biểu diễn chủ đề và phương pháp này ngăn ngừa những biến thể nổi bật trong sự phức tạp và năng lực trình diễn. Chúng bao gồm tần số, TF.IDF và các cách tiếp cận từ chủ đề bao gồm bảng các từ đơn và bộ trọng số tương ứng với thông tin là các từ có bộ trọng số càng cao thì càng biểu thị chủ đề.

Cách tiếp cận chuỗi từ vựng mà liệt kê từ liên quan đến lĩnh vực như WordNet được sử dụng để tìm các chủ đề hoặc khái niệm của những từ liên quan về ngữ nghĩa, và đưa ra trọng số cho các khái niệm. Phân tích ngữ nghĩa ẩn trong đó các mẫu từ đồng xuất hiện được xác định và phân tích đầy đủ như các chủ đề, tương tự như các trọng số cho mỗi mẫu.

Cách tiếp cận chủ đề Bayesian trong đó đầu vào được trình bày như sự hỗn độn các chủ đề và mỗi chủ đề đưa ra một bảng các phân phối xác suất từ (trọng số) cho chủ đề đó.

Các cách tiếp cận biểu diễn thuộc tính trình diễn mỗi câu trong đầu vào như là danh sách các thuộc tính quan trọng như là độ dài câu, vị trí trong văn bản, sự có mặt trong cụm,…

Trong các mô hình đồ thị, như là LexRank, toàn bộ văn bản được trình diễn như là mạng của các câu liên quan ngầm.

1.1.2. Ghi điểm các câu

Mỗi khi một sự biểu diễn trung gian được lấy ra, mỗi câu được ấn định một điểm số để xác định độ quan trọng. Với các cách tiếp cận biểu diễn chủ đề, điểm số thông thường liên quan tới độ phù hợp của một câu biểu thị một vài chủ đề quan trọng nhất trong văn bản hoặc mức độ nó kết nối thông tin xung quanh các chủ đề khác nhau. Với hầu hết các phương pháp biểu diễn thuộc tính, trọng số của câu được xác định bằng việc kết nối độ phù hợp từ các thuộc tính khác nhau, phổ biến nhất bằng việc sử dụng các kỹ thuật học máy để tìm ra bộ trọng số thuộc tính. Trong LexRank, trọng số của một câu được bắt nguồn từ việc áp dụng các kỹ thuật ngẫu nhiên tới sự biểu diễn đồ thị của văn bản.

1.1.3. Lựa chọn các câu tóm tắt

Cuối cùng, người tóm tắt phải lựa chọn việc kết nối tốt nhất các câu quan trọng để tạo ra một đoạn tóm tắt. Trong cách tiếp cận best n, nhóm n các câu quan trọng nhất được kết nối đã thỏa mãn chiều dài tóm tắt được lựa chọn cho bản tóm tắt. Trong cách tiếp cận tối đa hóa lề phù hợp, các câu được lựa chọn trong một thủ tục tham lam. Tại mỗi một bước của thủ tục, điểm số quan trọng của câu được tính lại như là một sự kết nối tuyến tính giữa trọng số quan trọng của câu và sự tương tự của nó với các câu vừa chọn. Các câu tương tự với các câu đã được lựa chọn sẽ bị loại bỏ. Trong cách tiếp cận lựa chọn toàn cục, sự thu thập tối ưu các câu là lựa chọn chủ đề liên quan tới các ràng buộc cố gắng làm cực đại hóa độ quan trọng toàn cục và cực tiểu hóa độ dư thừa và một số cách tiếp cận là cực đại hóa sự kết nối.

Có một vài ràng buộc giữa ba quá trình xử lý mô tả bên trên và một người tóm tắt có thể kết hợp bất kỳ sự kết nối các sự lựa chọn trong thực thi mỗi bước. Sự thay đổi trong phương pháp của mỗi bước cụ thể có thể thay đổi đáng kể tới chất lượng của bản tóm tắt. Trong việc sắp xếp độ quan trọng của việc tóm tắt, các nhân tố khác cũng được sử dụng. Nếu chúng ta có thông tin xung quanh ngữ cảnh để bản tóm tắt được sinh ra, điều này giúp xác định độ quan trọng. Ngữ cảnh có thể chứa các thông tin xung quanh nhu cầu người dùng, thường biểu thị thông qua một truy vấn. Ngữ cảnh có thể bao gồm môi trường trong đó một văn bản đầu vào được định vị như là các đường dẫn chỉ tới một trang web. Nhân tố khác ảnh hưởng tới sắp xếp câu là loại của văn bản. Khi văn bản đầu vào là một bản tin tức, một luồng email, một trang web hoặc một bài tạp chí ảnh hưởng tới chiến lược lựa chọn câu.

1.2. Tóm tắt tóm lược [22]

Tóm tắt tóm lược tạo ra một bản tóm tắt hiệu quả hơn so với tóm tắt trích chọn bởi việc nó có thể trích chọn thông tin từ tập các văn bản để khởi tạo bản tóm tắt thông tin rõ ràng. Một bản tóm tắt trình diễn thông tin tóm tắt trong một bản kết dính, dễ đọc và đúng ngữ pháp. Tính dễ đọc hay chất lượng ngữ pháp là một chất xúc tác để cải thiện chất lượng tóm tắt. Tóm tắt tóm lược được chia theo cách tiếp cận cấu trúc, theo cách tiếp cận ngữ nghĩa và gần đây là theo cách tiếp cận học sâu.

1.2.1. Cách tiếp cận cấu trúc

Cách tiếp cận cấu trúc mã hóa các thông tin quan trọng nhất trong văn bản thông qua kinh nghiệm như mẫu, các luật trích chọn và các cấu trúc khác như cây, ontology, lá và cấu trúc cụm.

1.2.1.1. Phương pháp cây

Kỹ thuật này sử dụng một cây phụ thuộc để biểu diễn văn bản, ngữ cảnh của một văn bản. Trong cách tiếp cận này các câu tương tự được tiền xử lý sử dụng một bộ phân tích cú pháp nông và sau đó câu được ánh xạ tới cấu trúc vị từ. Tiếp theo, bộ quản lý ngữ cảnh sử dụng thuật toán giao nhau để xác định cụm phổ biến bằng việc so sánh các cấu trúc vị từ. Các cụm này truyền các thông tin phổ biến được chọn lựa và sắp xếp cùng một số thông tin được thêm vào. Cuối cùng thuật toán sinh sử dụng ngôn ngữ sinh để kết nối và sắp xếp các cụm thành một câu tóm tắt mới. Điểm mạnh lớn nhất của cách tiếp cận này là sử dụng bộ sinh ngôn ngữ để cải thiện chất lượng tóm tắt tổng hợp như việc giảm thiểu sự lặp lại và tăng độ trôi chảy. Vấn đề gặp phải của cách tiếp cận này là ngữ cảnh sử dụng không bao gồm khi gặp cụm chồng chéo.

1.2.1.2. Phương pháp mẫu

Kỹ thuật này sử dụng một mẫu để biểu diễn toàn bộ văn bản. Các mẫu ngôn ngữ hay luật trích chọn được so khớp để xác định các mảnh văn bản, được ánh xạ tới các vị trí mẫu. Các mẫu văn bản là thể hiện của ngữ cảnh tóm tắt.

1.2.1.3. Phương pháp Ontology

Phương pháp này được sử dụng để nâng cao chất lượng tóm tắt. Miền ontology cho các sự kiện tin tức được xác định bởi các chuyên gia. Pha tiếp theo là pha xử lý văn bản. Các mục có nghĩa từ tập văn bản được sinh ra trong pha này. Các mục có nghĩa được phân loại bằng người dựa trên các khái niệm của sự kiện tin tức. Giới hạn của cách tiếp cận là thời gian rảnh bởi vì miền ontology được xác định bởi chuyên gia.

1.2.1.4. Phương pháp luật

Phương pháp này bao gồm ba bước. Đầu tiên, văn bản được phân loại để biểu diễn các hạng mục của các nhóm. Các nhóm có thể đến từ các miền khác nhau. Bước tiếp theo là phần câu hỏi trên các nhóm. Ví dụ các nhóm như chiến tranh, bệnh tật, sức khỏe,.. lấy ví dụ các câu hỏi trong nhóm như: điều gì xảy ra?, khi nào xảy ra?, ai ảnh hưởng tới?, hậu quả là gì?… Phụ thuộc vào các câu hỏi này các luật được sinh ra. Ở đây một vài động từ và danh từ có nghĩa tương tự được xác định và vị trí của chúng được xác định đúng. Mô hình lựa chọn ngữ cảnh đưa ra ứng cử tốt nhất trong tổng số đó. Bộ sinh mẫu được sử dụng cho việc sinh câu tóm tắt.

1.2.2. Cách tiếp cận ngữ nghĩa

Trong cách tiếp cận ngữ nghĩa, biểu diễn ngữ nghĩa của văn bản được sử dụng để cung cấp cho hệ thống sinh ngôn ngữ. Cách tiếp cận này tập trung vào xác định các cụm danh từ và cụm động từ.

1.2.2.1. Mô hình ngữ nghĩa đa phương thức

Trong cách tiếp cận này, một mô hình ngữ nghĩa thu thập các khái niệm và quan hệ giữa các khái niệm, được xây dựng để biểu thị ngữ cảnh của tập các văn bản. Khái niệm quan trọng được định vị dựa trên một vài độ đo và các khái niệm cuối cùng được trình bày như các câu trong bản tóm tắt.

1.2.2.2. Phương thức dựa trên thông tin

Trong cách tiếp cận này, các khái niệm của bản tóm tắt được sinh ra từ sự biểu diễn trừu tượng của văn bản nguồn, hơn là từ các câu của văn bản nguồn. Biểu diễn trừu tượng là thành phần quan trọng nhất của thông tin kết dính trong văn bản. Từ phương pháp này, một thông tin ngắn gọn, kết dính được làm giàu và bản tóm tắt giảm dư thừa được hình thành. Mặc dù chứa nhiều thuận lợi, phương pháp này cũng có những giới hạn. Trong khi đưa ra các câu đúng ngữ pháp và có nghĩa, nhiều thông tin quan trọng bị bỏ qua.

1.2.2.3. Phương pháp dựa trên đồ thị ngữ nghĩa

Phương pháp này nhắm tới việc tóm tắt bằng việc khởi tạo một đồ thị ngữ nghĩa gọi là Đồ thị ngữ nghĩa giàu (RSG) cho văn bản gốc, giảm thiểu các đồ thị ngữ nghĩa sinh ra, và sau đó sinh ra bản tóm tắt trừu. Cách tiếp cận gồm ba giai đoạn. Đầu tiên, cụm biểu diễn văn bản đầu vào sử dụng đồ thị ngữ pháp, động từ và danh từ của văn bản đầu vào được biểu diễn như là các nút đồ thị và các cạnh tương thích với quan hệ ngữ nghĩa và hình topo giữa chúng. Giai đoạn thứ hai giảm thiểu đồ thị ban đầu tới nhiều đồ thị sử dụng luật thông minh. Điểm thuận lợi của phương pháp này là giảm thiểu thông tin dư thừa và đưa ra câu đúng ngữ pháp. Điểm bất lợi của phương pháp là sự giới hạn tới một văn bản mà không cho đa văn bản.

1.2.3. Cách tiếp cận học sâu

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của các mô hình huấn luyện end-to-end đã tạo ra hướng đi mới để giải quyết bài toán tóm tắt văn bản tự động. Mặc dù vậy tóm tắt tóm lược ý sử dụng học sâu vẫn đang ở trong giai đoạn đầu phát triển. Bản tóm tắt tạo ra còn chưa đúng ngữ pháp, nhiều từ dư thừa và không chứa đủ thông tin quan trọng của văn bản.

Do đó, tôi muốn áp dụng các phương pháp học sâu hiện đại vào bài toán tóm tắt văn bản theo hướng tóm lược ý, nhằm mục đích cải thiện chất lượng tóm tắt văn bản và đồng thời đưa ra một mô hình mạnh mẽ cho bài toán này.

Chương 2: Cơ sở lý thuyết

Những nghiên cứu đầu tiên cho bài toán tóm tắt văn bản theo phương pháp mạng nơ-ron thuộc về nhóm tác giả Alexander M. Rush [2]. Họ ước lượng một mô hình attention cục bộ, đưa ra một từ của bản tóm tắt dựa theo câu đầu vào. Nghiên cứu dựa trên sự phát triển của các phương pháp dịch máy nơ-ron. Họ kết hợp mô hình xác suất với một thuật toán sinh để đưa ra độ chính xác của tóm tắt. Mặc dù mô hình đơn giản về cấu trúc nhưng có thể dễ dàng được huấn luyện end-to-end và mở rộng với một số lượng dữ liệu huấn luyện lớn hơn. Ngay sau đó, Submit Chorpa cùng cộng sự [3] giới thiệu một mạng truy hồi RNN có điều kiện để đưa ra một tóm tắt. Ràng buộc điều kiện được cung cấp bởi mạng xoắn convolution attention encoder đảm bảo bộ giải mã tập trung ở các từ đầu vào phù hợp tại mỗi bước. Mô hình dựa vào khả năng học các đặc trưng và dễ dàng học end-to-end trên một lượng lớn dữ liệu. Cùng với đó, nhóm của Ramesh Nallapti [19] đưa ra bản tóm tắt sử dụng mạng RNN Attention Encoder-Decoder. Kết quả đạt cao nhất trên hai bộ dữ liệu khác nhau.

Dưới đây tôi xin trình bày những khái niệm và mô hình cơ bản trong lý thuyết mạng nơ-ron.

2.1. Mạng nơ-ron [21]

Phần này cung cấp một cái nhìn tổng quan về mạng nơ-ron nhân tạo, với sự nhấn mạnh vào ứng dụng vào các nhiệm vụ phân loại và ghi nhãn.

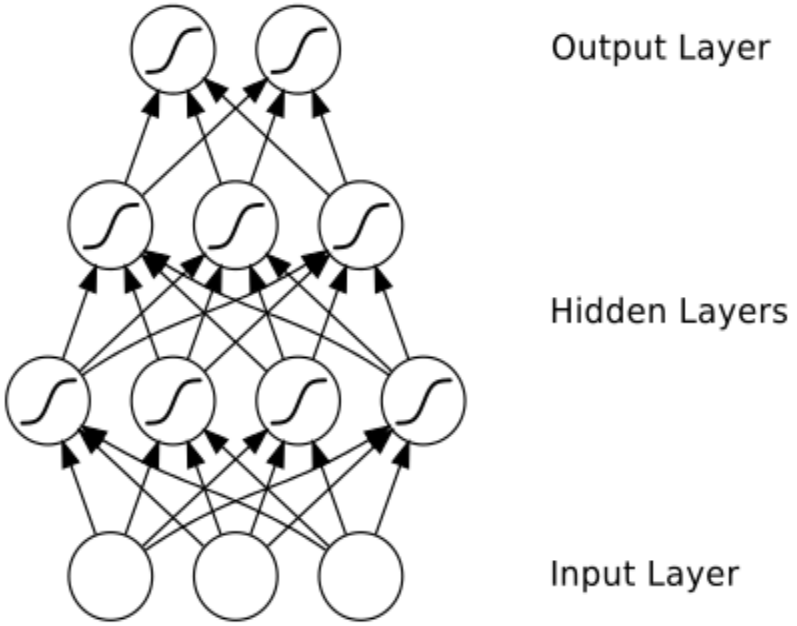
2.1.1. Mạng nơ-ron đa lớp (Multilayer Perceptrons)

Mạng nơ-ron nhân tạo (ANNs) đã được phát triển như là mô hình toán học bằng năng lực xử lý thông tin của bộ não sinh học (McCulloch và Pitts, 1988; Rosenblatt, 1963; Rumelhart et al., 1986).

Cấu trúc cơ bản của một ANN là một mạng lưới các tế bào nhỏ, hoặc nút, tham gia với nhau bởi các kết nối trọng số. Xét về mặt mô hình sinh học gốc, các nút đại diện cho tế bào nơ-ron, và các trọng số kết nối đại diện cho sức mạnh của các khớp nơ-ron giữa các tế bào nơ-ron. Các mạng kích hoạt bằng cách cung cấp một đầu vào cho một số hoặc tất cả các nút, và kích hoạt này sau đó lây lan khắp các mạng cùng các kết nối trọng số.

Nhiều biến thể của mạng ANNs đã xuất hiện trong những năm qua, với tính chất rất khác nhau . Một khác biệt quan trọng giữa ANNs là kết nối dạng chu kỳ và những kết nối khác dạng mạch hở. ANNs với chu kỳ được gọi là mạng nơ-ron

phản hồi đệ quy. Mạng ANN không có chu trình được gọi là mạng lan truyền tiến (FNNs). Ví dụ nổi tiếng của FNNs bao gồm perceptron (Rosenblatt, 1958), mạng hàm cơ sở xuyên tâm (Broomhead và Lowe, 1988), bản đồ Kohonen (Kohonen, 1989) và Hopfield lưới (Hopfield, 1982). Các hình thức sử dụng rộng rãi nhất của FNN và những gì ta tập trung vào trong phần này, là Perceptron đa lớp (MLP, Rumelhart et al, 1986; Werbos, 1988; Bishop, 1995).



Alex Graves [21]

Hình 2.1: Một perceptron nhiều lớp.

Như minh họa trong hình 2.1, các đơn vị trong một Perceptron đa lớp được bố trí trong lớp, với các kết nối lan truyền tới một lớp kế tiếp. Mô hình được bắt nguồn từ các lớp đầu vào, sau đó truyền qua lớp ẩn đến lớp ra. Quá trình này được gọi là lan truyền về phía trước của mạng.

Do đầu ra của một MLP chỉ phụ thuộc vào đầu vào hiện tại, và không trên bất kỳ đầu vào từ quá khứ hay tương lai, MLPs phù hợp hơn cho mô hình phân loại hơn so với ghi nhãn theo thứ tự.

Một MLP chứa một tập hợp các giá trị trọng số định nghĩa một hàm ánh xạ vector đầu vào tới vector đầu ra. Bằng cách thay đổi trọng số, một MLP duy nhất có khả năng đại diện cho nhiều hàm khác nhau. Thực tế nó đã được chứng minh

(Hornik et al., 1989) rằng một MLP với một lớp ẩn chứa một số lượng đủ các đơn vị không tuyến tính có thể xấp xỉ hàm liên tục trên một tên miền đầu vào đến độ chính xác tùy ý. Vì lý do này MLPs được cho là hàm xấp xỉ tổng quát.

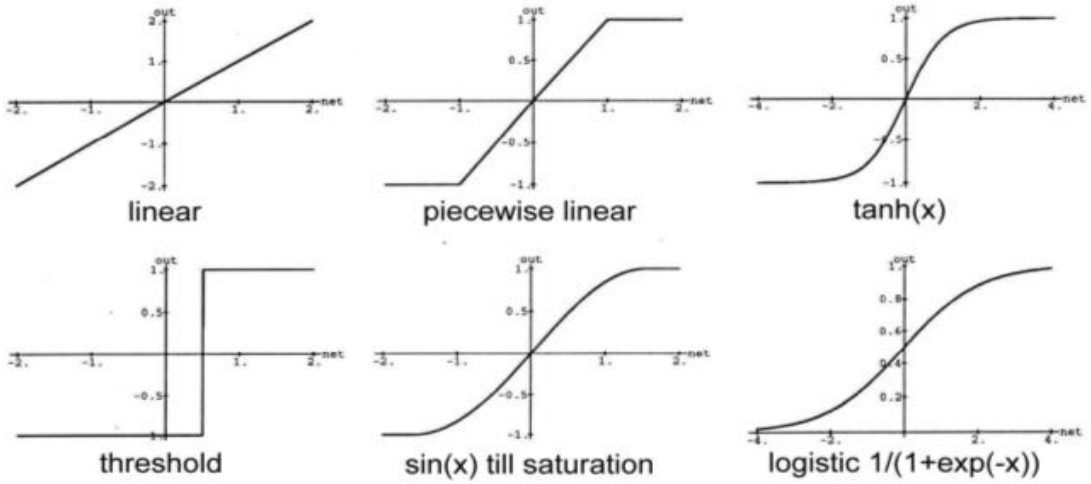
2.1.2. Lan truyền tiến (Forward Pass)

Hãy xem xét một MLP với I đơn vị đầu vào, kích hoạt bằng vector đầu vào x (| x | = I). Mỗi đơn vị trong lớp ẩn đầu tiên sẽ tính tổng trọng số của các đơn vị đầu vào. Đối với đơn vị ẩn h, được đề cập là đầu vào mạng tới đơn vị h, và biểu thị nó là a_h. Sau đó các hàm kích hoạt θ_h được áp dụng, đưa ra kết quả b_h của đơn vị. Biểu thị trọng số từ đơn vị i tới đơn vị j như w_{ij}, ta có:

$$a_h = \sum_{i=1}^{I} w_{ih} x_i$$

$$b_h = \theta_h(a_h) \tag{2.1}$$

Một số hàm kích hoạt hệ thống nơ-ron được vẽ trong hình bên dưới, phổ biến nhất là hàm tanh hyperbol.



Alex Graves [21]

Hình 2.2: Hàm kích hoạt mạng nơ-ron.

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{2.2}$$

Hàm hàm sigmoid:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

Hai hàm có liên quan bởi các biến đổi tuyến tính sau đây:

$$\tanh(x) = 2\sigma(2x) - 1 \tag{2.4}$$

Điều này có nghĩa rằng bất kỳ hàm tính toán bởi một mạng lưới nơ-ron với một lớp ẩn của đơn vị tanh có thể được tính toán bởi một mạng lưới với các đơn vị logistic sigmoid và ngược lại. Do đó, các hàm kích hoạt là tương đương. Tuy nhiên một lý do để phân biệt giữa chúng là dãy đầu ra của chúng là khác nhau; đặc biệt nếu một đầu ra giữa 0 và 1 được yêu cầu (ví dụ, nếu kết quả đại diện xác suất) thì hàm sigmoid nên được sử dụng.

Một điểm quan trọng của cả hai hàm tanh và hàm sigmoid là tính không tuyến tính của chúng. Mạng nơron phi tuyến là mạnh hơn so với các mạng tuyến tính tương ứng. Hơn nữa, bất kỳ sự kết hợp của thao tác tuyến tính chính là một thao tác tuyến tính, có nghĩa là bất kỳ MLP với nhiều lớp tuyến tính ẩn là tương đương với một số MLP khác với một lớp ẩn đơn tuyến tính. Do đó, các mạng phi tuyến có thể đạt được sức mạnh đáng kể bằng cách sử dụng liên tiếp lớp ẩn để đại diện cho dữ liệu đầu vào (Hinton et al, 2006; Bengio và LeCun, 2007).

Một điểm quan trọng là cả hai hàm này là khả vi, cho phép mạng được huấn luyện với gradient descent. Các dẫn xuất đầu tiên của chúng là:

$$\frac{\partial \tanh(x)}{\partial x} = 1 - \tanh(x)^2$$
$$\frac{\partial \tanh(x)}{\partial x} = \sigma(x)(1 - \sigma(x)) \tag{2.5}$$

Do cách chúng làm giảm một miền đầu vào vô hạn với một loạt phạm vi đầu ra hữu hạn, hàm kích hoạt mạng lưới nơ-ron đôi khi được gọi là hàm ép.

Sau khi tính toán các kích hoạt của các đơn vị trong lớp ẩn đầu tiên, quá trình tổng hợp và kích hoạt được sau đó lặp lại đối với phần còn lại của các lớp ẩn theo thứ tự lần lượt, ví dụ cho đơn vị h trong tâng ẩn thứ l H_l

$$a_h = \sum_{h' \in H_{l-1}} w_{h'h} b_{h'}$$

$$b_h = \theta_h(a_h)$$

(2.6)

2.1.3. Tầng đầu ra (Output Layers)

Các vector đầu ra y của một MLP được đưa ra bởi sự kích hoạt của các đơn vị trong lớp ra. Các mạng đầu vào a_k cho mỗi đơn vị đầu ra k được tính bằng tổng các đơn vị kết nối với nó, chính xác cho một đơn vị ẩn. Điều này đúng cho một mạng L lớp ẩn.

$$a_k = \sum_{h \in H_L} w_{hk} b_h$$

(2.7)

Việc chọn số đơn vị trong tầng đầu ra và lựa chọn hàm kích hoạt đầu ra phụ thuộc vào các nhiệm vụ mạng áp dụng. Đối với nhiệm vụ phân loại nhị phân, cấu hình tiêu chuẩn là đơn vị duy nhất với một hàm kích hoạt sigmoid. Vì phạm vi của các sigmoid logistic là khoảng mở (0, 1), sự kích hoạt của các đơn vị đầu ra có thể được giải thích như là xác suất mà các vector đầu vào thuộc lớp đầu tiên (và ngược lại, một trừ đi kích hoạt cho các xác suất mà nó thuộc về lớp thứ hai).

$$p(C_1|x) = y = \sigma(a)$$

$$p(C_2|x) = 1 - y$$

(2.8)

Việc sử dụng các hàm sigmoid là một ước lượng xác suất nhị phân đôi khi gọi là hồi quy logistic, hoặc một mô hình logit. Nếu chúng ta sử dụng một chương trình mã hóa cho vector mục tiêu z nơi z = 1 nếu lớp đúng là C1 và z = 0 nếu đúng lớp học là C2, chúng ta có thể kết hợp các biểu thức trên để viết:

$$p(z|x) = y^z(1-y)^{1-z}$$

(2.9)

Đối với vấn đề phân loại với K> 2 lớp, quy ước là có K đơn vị đầu ra, và chuẩn hóa kích hoạt đầu ra với các hàm softmax (Bridle, 1990) để có được các xác suất lớp:

$$p(C_k|x) = y_k = \frac{e^{a_k}}{\sum_{k'=1}^{K} e^{a_{k'}}}$$

(2.10)

Đây còn được biết đến như là một mô hình đa logit. Một lược đồ 1-of-K giới thiệu về lớp mục tiêu z là một vector nhị phân với tất cả các yếu tố bằng số không, trừ cho các phần tử tương ứng lớp đúng bằng với một. Ví dụ, nếu K = 5 và lớp đúng là C2, z được đại diện bởi (0, 1, 0, 0, 0).

Chúng ta có được xác suất mục tiêu:

$$p(z|x) = \prod_{k=1}^{K} y_k^{z_k}$$

(2.11)

Với các định nghĩa trên, việc sử dụng MLPs cho mô hình phân loại là đơn giản. Chỉ cần đi trong một vector đầu vào, kích hoạt mạng, và chọn nhãn lớp tương ứng với đơn vị đầu ra tích cực nhất.

2.1.4. Hàm lỗi (Loss Functions)

Đối với phân loại nhị phân, thay thế (2.9) vào tối đa độ phù hợp hàm lỗi:

$$L(x, z) = - \ln p(z | x)$$

Ta có:

$$\mathcal{L}(\mathrm{x}, \mathrm{z}) = (z-1)\ln(1-y) - z\ln y$$

(2.12)

Tương tự như vậy, đối với vấn đề với nhiều lớp học,

$$\mathcal{L}(x, z) = -\sum_{k=1}^{K} z_k \ln y_k$$

(2.13)

2.1.5. Lan truyền ngược (Backward Pass)

Kể từ MLPs, bằng cách xây dựng, khai thác khả vi, chúng có thể được huấn luyện để giảm thiểu bất kỳ chức năng mất khả vi sử dụng gradient descent. Ý tưởng cơ bản của gradient descent là tìm đạo hàm của hàm lỗi đối với cho mỗi trọng số mạng, sau đó điều chỉnh các trọng số theo hướng độ dốc âm.

Để tính toán hiệu quả gradient, ta sử dụng một kỹ thuật gọi là lan truyền ngược (Rumelhart et al, 1986;. Williams và Zipser, 1995; Werbos, 1988). Điều này thường được gọi là các đường chuyền quay lui của hệ thống mạng. Lan truyền ngược đơn giản chỉ là một ứng dụng lặp đi lặp lại các quy tắc dây chuyền cho một phần các dẫn xuất. Bước đầu tiên là để tính toán các đạo hàm của hàm lỗi với đối với các đơn vị đầu ra. Đối với một mạng lưới phân loại nhị phân, đạo hàm hàm lỗi được xác định trong (2.12) đối với các kết quả đầu ra mạng cho

$$\frac{\partial \mathcal{L}(x, z)}{\partial y} = \frac{y - z}{y(1 - y)}$$

(2.14)

Các quy tắc dây chuyền cho chúng ta:

$$\frac{\partial \mathcal{L}(x, z)}{\partial a} = \frac{\partial \mathcal{L}(x, z)}{\partial y} \frac{\partial y}{\partial a}$$

(2.15)

và sau đó chúng ta có thể thay thế để có được

$$\frac{\partial \mathcal{L}(x, z)}{\partial a} = y - z$$

(2.16)

Đối với một mạng lưới nhiều lớp,

$$\frac{\partial \mathcal{L}(x, z)}{\partial y_k} = -\frac{z_k}{y_k}$$

(2.17)

Sự kích hoạt của mỗi đơn vị trong một lớp softmax phụ thuộc vào mạng đầu vào cho từng đơn vị trong lớp, quy tắc dây chuyền cung cấp cho ta công thức:

$$\frac{\partial \mathcal{L}(x,z)}{\partial a_k} = \sum_{k'=1}^{K} \frac{\partial \mathcal{L}(x,z)}{\partial y_{k'}} \frac{\partial y_{k'}}{\partial a_k}$$

$$\frac{\partial y_{k'}}{\partial a_k} = y_k \delta_{kk'} - y_k y_{k'}$$

(2.18)

và sau đó chúng ta có được:

$$\frac{\partial \mathcal{L}(x,z)}{\partial a_k} = y_k - z_k$$

(2.19)

Bây giờ chúng ta tiếp tục áp dụng các quy tắc dây chuyền, làm ngược qua lớp ẩn.

Tại thời điểm này, ta đặt ký hiệu sau:

$$\delta_j \stackrel{\text{def}}{=} \frac{\partial \mathcal{L}(x,z)}{\partial a_j}$$

(2.20)

trong đó j là bất kỳ đơn vị trong mạng. Đối với các đơn vị trong lớp ẩn cuối cùng, ta có

$$\delta_h = \frac{\partial \mathcal{L}(x,z)}{\partial b_h} \frac{\partial b_h}{\partial a_h} = \frac{\partial b_h}{\partial a_h} \sum_{k=1}^{K} \frac{\partial \mathcal{L}(x,z)}{\partial a_k} \frac{\partial a_k}{\partial b_h}$$

(2.21)

nơi mà ta đã sử dụng thực tế là L(x, z) chỉ phụ thuộc vào mỗi đơn vị h ẩn thông qua ảnh hưởng của nó đối với các đơn vị đầu ra.

$$\delta_h = \theta'(a_j) \sum_{k=1}^{K} \delta_k w_{hk}$$

(2.22)

Các giá trị δ cho mỗi lớp ẩn H_l trước khi cuối cùng có thể được tính đệ quy:

$$\delta_h = \theta'(a_h) \sum_{h' \epsilon H_{l+1}} \delta_{h'} w_{hh'}$$

(2.23)

Một khi chúng ta có những giá trị δ cho tất cả các đơn vị ẩn, chúng ta có thể sử dụng để tính toán các đạo hàm đối với mỗi trọng số:

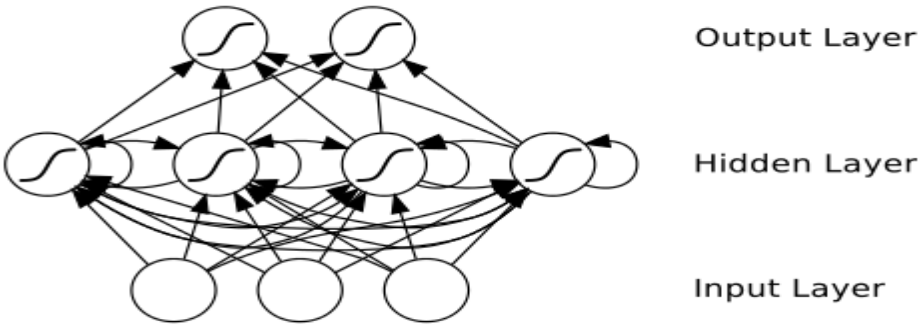
$$\frac{\partial \mathcal{L}(x,z)}{\partial w_{ij}} = \frac{\partial \mathcal{L}(x,z)}{\partial a_j} \frac{\partial a_j}{\partial w_{ij}} = \delta_j b_i$$

(2.24)

2.2. Mô hình RNN

Ở phần trước ta đã xem xét mạng nơ-ron hướng tiến mà các kết nối không tạo thành chu kỳ. Nếu ta giả định điều kiện này, cho phép kết nối theo chu kỳ, chúng ta sẽ đạt được mạng nơ-ron hồi quy (Recurrent Neural Network - RNN).

Điểm khác biệt giữa một mạng nơ-ron đa tầng và một mạng nơ-ron hồi quy có vẻ đơn giản, ngụ ý việc học chuỗi được tiếp cận sâu rộng hơn. Một mạng MLP chỉ có thể ánh xạ từ đầu vào tới các vector đầu ra, ngược lại RNN có thể ánh xạ bắt nguồn từ toàn bộ lịch sử của các đầu vào đằng trước tới mỗi đầu ra. Tổng quát hơn, tương đương kết quả cho MLP là một RNN với một số lượng đủ các đơn vị ẩn có thể ước tính bất kỳ chuỗi tới độ chính xác tùy ý. Điểm mấu chốt ở đây là các kết nối hồi quy cho phép một bộ nhớ của các tầng đầu vào đằng trước tồn tại bên trong trạng thái của mạng và do đó ảnh hưởng tới đầu ra mạng.



Alex Graves [21]

Hình 2.3: Một mạng RNN

2.2.1. Pha hướng tiến

Pha hướng tiến của một RNN giống với một mạng nơ-ron đa tầng với một tầng ẩn, trừ việc hàm kích hoạt đến từ tầng ẩn của cả đầu vào bên ngoài hiện tại và các hàm kích hoạt tầng ẩn từ trạng thái đằng trước. Xem xét một đầu vào x độ dài T tới mạng RNN với I đơn vị đầu vào, H đơn vị ẩn, và K đơn vị đầu ra. Cho phép x_i^t là giá trị của đầu vào i tại thời điểm t, và a_j^t và b_j^t tương ứng với đầu vào mạng tới đơn vị j tại thời điểm t và hàm kích hoạt của đơn vị j tại thời điểm t. Đối với các đơn vị ẩn, chúng ta có:

$$a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t-1} \tag{2.25}$$

Để phi tuyến, các hàm kích hoạt khác nhau được áp dụng chính xác như một mạng MLP:

$$b_h^t = \theta_h(a_h^t) \tag{2.26}$$

Hoàn tất một chuỗi các hàm kích hoạt ẩn có thể được tính toán bắt đầu tại t=1 và được áp dụng hồi quy, tăng dần t tại mỗi thời điểm. Chú ý rằng cần khởi tạo giá trị b_i^0 để chọn các đơn vị ẩn, tương ứng với các trạng thái mạng trước khi nó nhận bất kỳ thông tin từ chuỗi dữ liệu.

Các đầu vào mạng tới các đơn vị đầu ra có thể được tính toán tại cùng thời điểm với các hàm kích hoạt ẩn:

$$a_k^t = \sum_{h=1}^{H} w_{hk} b_h^t \tag{2.27}$$

2.2.2. Pha quay lui

Cho đạo hàm từng phần một số hàm lỗi L với các đầu ra mạng tương ứng và tiếp theo là xác định các đạo hàm với các trọng số tương ứng. Thuật toán quay lui thường được áp dụng cho mạng RNN vì tính đơn giản và hiệu quả về thời gian tính toán.

Giống như thuật toán quay lui chuẩn, thuật toán lặp lại các quy tắc chuỗi. Sự tinh tế ở chỗ, đối với mạng hồi quy, hàm lỗi phụ thuộc vào sự kích hoạt tầng ẩn không chỉ ảnh hưởng trên lớp đầu ra mà còn thông qua ảnh hưởng trên tầng ẩn tại thời điểm tiếp theo. Vì thế:

$$\delta_h^t = \theta'(a_h^t)(\sum_{k=1}^{K} \delta_k^t w_{hk} + \sum_{h'=1}^{H} \delta_{h'}^{t+1} w_{hh'})$$

(2.28)

Trong đó:

$$\delta_j^t \stackrel{\text{def}}{=} \frac{\partial \mathcal{L}}{\partial a_j^t}$$

(2.29)

Chuỗi hoàn tất δ có thể được tính toán bắt đầu từ t=T và áp dụng hồi quy, giảm bớt t tại mỗi bước. Chú ý rằng $\delta_j^{T+1} = 0$ với mọi j khi không có lỗi từ ngoài phần cuối của chuỗi. Cuối cùng, trọng số tương tự được tái sử dụng tại mọi thời điểm, ta tổng hợp lại toàn bộ chuỗi để nhận được đạo hàm tương ứng với bộ trọng số của mạng:

$$\frac{\partial \mathcal{L}}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{ij}} = \sum_{t=1}^{T} \delta_j^t b_i^t$$

(2.30)

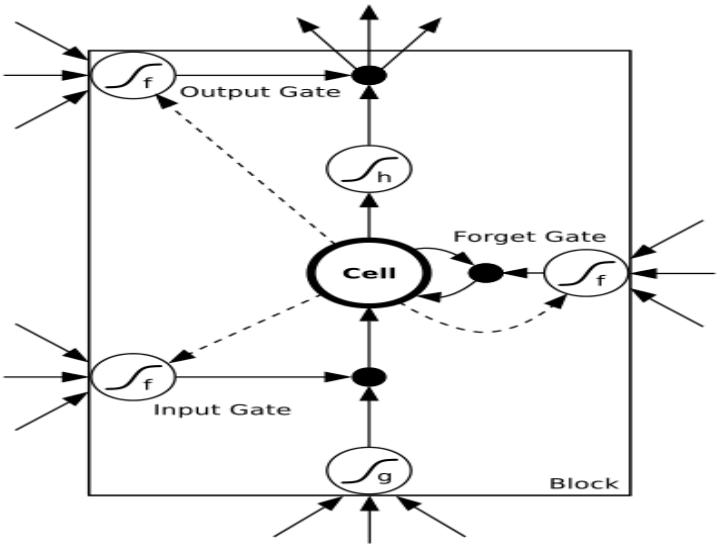
Khi huấn luyện RNN, ta sử dụng kỹ thuật đạo hàm quay lui, để cộng dồn đạo hàm của các bước quay lại với nhau. Đây là một biện pháp để giải quyết vấn đề đạo hàm hội tụ về 0 qua các bước lặp nhưng cũng cần điều chỉnh phù hợp để đạo hàm không phân kỳ. Đó cũng là vấn đề đặt ra trong nhiều năm và mạng LSTM (Hochreiter & Schmidhuber - 1997) và mới đây là mạng GRU (Cho - 2014) được đề xuất để giải quyết vấn đề này.

2.3. Mạng LSTM, GRU

2.3.1. Mạng LSTM

Như đã trình bày phần trước, một điểm thuận lợi của mạng nơ-ron hồi quy là khả năng sử dụng thông tin ngữ cảnh khi ánh xạ giữa chuỗi đầu vào và chuỗi đầu ra. Tuy nhiên, với kiến trúc RNN tiêu chuẩn phạm vi của ngữ cảnh có thể được truy cập khá hạn chế. Vấn đề là do ảnh hưởng của đầu vào trên tầng ẩn, và vì thế trên đầu ra của mạng hoặc là suy giảm hoặc là tăng lên cấp số nhân theo chu kỳ xung quanh các kết nối hồi quy của mạng. Hiệu ứng này còn gọi là vấn đề biến mất đạo hàm (vanishing gradient problem). Một lượng lớn các nghiên cứu được đưa ra vào những năm 1990 để giải quyết vấn đề giảm đạo hàm cho mạng RNN. Các nghiên cứu bao gồm quá trình huấn luyện không cần tính đạo hàm, như thuật toán giả mô phỏng và rời rạc lỗi truyền, hoặc dùng thời gian trễ, thời gian ràng buộc. Mạng LSTM (Long Short Term Memory) được đưa ra là cũng cách tiếp cận giải quyết vấn đề này.

Kiến trúc mạng LSTM bao gồm một tập các mạng con được kết nối hồi quy, còn gọi là các khối nhớ. Các khối có thể được liên tưởng như là phiên bản khác của các chip nhớ trong máy tính số. Mỗi khối nhớ chứa một hoặc nhiều ô nhớ tự liên kết và ba đơn vị: đầu vào, đầu ra và cổng quên cung cấp khả năng liên tục viết, đọc và hoạt động khởi động cho các ô nhớ.



Alex Graves [21]

Hình 2.4: Một khối nhớ LSTM với một ô nhớ

Một mạng LSTM tương đương với mạng RNN trừ việc các đơn vị tổng hợp trong tầng ẩn được thay thế bằng các khối nhớ. Các khối LSTM cũng có thể được hòa trộn với các đơn vị tổng hợp mặc dù về cơ bản là không cần thiết. Tầng đầu ra có thể được sử dụng cho các mạng LSTM như cho mạng RNN chuẩn.

Các cổng nhân lên cho phép các ô nhớ LSTM được lưu trữ và truy cập thông tin trên một thời gian dài, vì thế giảm nhẹ vấn đề biến mất đạo hàm. Ví dụ ngay khi cổng đầu vào được đóng lại (có hàm kích hoạt gần 0), sự kích hoạt của ô sẽ không bị ghi đè bởi đầu vào đang đến trong mạng, do đó có thể cung cấp cho mạng sau này bằng cách mở cổng đầu ra.

LSTM khá thành công trong một loạt các nhiệm vụ yêu cầu bộ nhớ phạm vi dài, và nó còn được áp dụng trong các vấn đề trong thế giới thực như là cấu trúc thứ cấp proteion, sinh âm nhạc, nhận dạng âm thanh, nhận dạng chữ viết.

2.3.2. Mạng GRU

Mạng RNN làm việc trên biến tuần tự x = (x_1, x_2,…,x_T) bởi việc duy trì trạng thái ẩn h quá thời gian. Tại mỗi thời điểm t, trạng thái ẩn h được cập nhật bằng công thức:

$$h^{(t)} = f(h^{(t-1)}, x_t)$$

(2.31)

Trong đó: f là hàm kích hoạt. Thông thường f thực thi như là một hàm chuyển tuyến tính trên vector đầu vào, tổng hợp lại thành một hàm logistic sigmoid.

RNN được sử dụng hiệu quả cho việc học phân phối các biến tuần tự bằng việc học phân phối trên đầu vào p(x_{t+1}|x_t,…,x_1). Ví dụ, trong trường hợp chuỗi 1 đến K vector, phân phối có thể học bởi một mạng RNN, đưa ra đầu ra:

$$p(x_{t,j} = 1|x_{t-1}, \dots, x_1) = \frac{\exp(w_j h_{(t)})}{\sum_{j'=1}^{K} \exp(w_{j'} h_{(t)})}$$

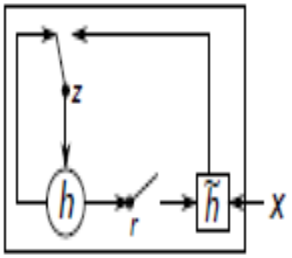
(2.32)

Cho tất cả các giá trị j = 1,..,K. Trong đó, w_j là tất cả các hàng của ma trận trọng số W. Kết quả trong phân phối:

$$p(x) = \prod_{t=1}^{T} p(x_t | x_{t-1}, \dots, x_1)$$

(2.33)

Gần đây, Cho[12] giới thiệu mạng GRU (Gated Recurrent Units) như là một mạng kích hoạt cho RNN. Hàm kích hoạt mới làm tăng thêm hàm kích hoạt sigmoid với hai cổng gọi là reset r, và update z. Mỗi cổng phụ thuộc vào trạng thái ẩn đằng trước $h^{(t-1)}$ và đầu vào hiện tại x_t đưa ra luồng thông tin.



Kyunghyun Cho et al. [12]

Hình 2.5: Minh họa mạng GRU

Đầu tiên cổng reset r_j được tính toán như sau:

$$r_j = \sigma([W_r x]_j + [U_r h_{(t-1)}]_j)$$

(2.34)

Trong đó: σ là làm kích hoạt logistic sigmoid

$[.]_j$ xác định thành phần thứ j của vector, x và h_{t-1} là đầu vào và trạng thái ẩn đằng trước tương ứng. W_r và U_r là ma trận trọng số cần học.

Tương tự cổng update z được tính bằng:

$$z_j = \sigma([W_z x]_j + [U_z h_{(t-1)}]_j)$$

(2.35)

Trạng thái ẩn h_j được tính bằng công thức:

$$h_j^{(t)} = z_j h_j^{(t-1)} + (1 - z_j)\tilde{h}_j^{(t)}$$

(2.36)

Trong đó:

$$\tilde{h}_j^{(t)} = \phi([Wx]_j + [U(r \odot h_{t-1})]_j)$$

(2.37)

Khi cổng reset tiến gần tới 0, trạng thái ẩn dần bỏ qua sự có mặt của trạng thái ẩn đằng trước và chỉ ảnh hưởng bởi đầu vào hiện tại. Điều này cho phép trạng thái ẩn hủy bỏ bất kỳ thông tin nào không phù hợp trong tương lai, cho phép trình diễn gọn nhẹ hơn.

Mặt khác, cổng update điều khiển việc bao nhiêu thông tin từ trạng thái ẩn đằng trước được mang tới trạng thái ẩn hiện tại. Điều này giúp RNN nhớ thông tin lâu hơn.

2.4. Mạng nơ-ron tích chập

Mạng nơ-ron tích chập (Convolution Neural Network - CNN – LeCun, 1989) là một mạng nơ-ron cho xử lý dữ liệu dạng lưới. CNN đã áp dụng khá thành công trong các ứng dụng như xử lý ảnh, xử lý tiếng nói, xử lý âm thanh,… Tên gọi mạng nơ-ron tích chập có nghĩa là mạng sử dụng một biểu thức toán học gọi là tích chập. Tích chập là một dạng đặc biệt của phép tuyến tính. Như vậy mạng CNN là một mạng nơ-ron đơn giản sử dụng phép tích chập trong các phép nhân ma trận tại ít nhất một trong các tầng của nó.

Phép tích chập có bắt nguồn trong xử lý ảnh. Để làm mịn ảnh có nhiễu, người ta sử dụng trung bình một vài độ đo. Gọi x(t) là giá trị điểm ảnh tại vị trí t. Gọi w(a) là hàm trọng số, trong đó a là đại diện cho độ đo.

Nếu chúng ta áp dụng phép lấy trung bình bộ trọng số tại mọi thời điểm. ta sẽ đạt được hàm mịn s tại vị trí nhiễu.

$$s(t) = \int x(a)w(t-a)da$$

(2.38)

Đây được gọi là phép tích chập. Phép tích chập thường được xác định bằng dấu *

$$s(t) = (x * w)(t)$$

(2.39)

Trong hệ thống mạng nơ-ron tích chập, tham số đầu tiên x được xác định như đầu vào và tham số thứ hai w được xác định như hàm nhân. Đầu ra được xác định như là ánh xạ đặc trưng.

Khi làm việc với dữ liệu trên máy tính, thời gian là rời rạc, được nhận những giá trị kiểu số. Khi đó dạng rời rạc của mạng nơ-ron tích chập là:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a)$$

(2.40)

Trong các ứng dụng học máy, đầu vào luôn luôn là mảng dữ liệu đa chiều và nhân luôn là một mảng các tham số đa chiều có tác dụng điều chỉnh thuật toán học.

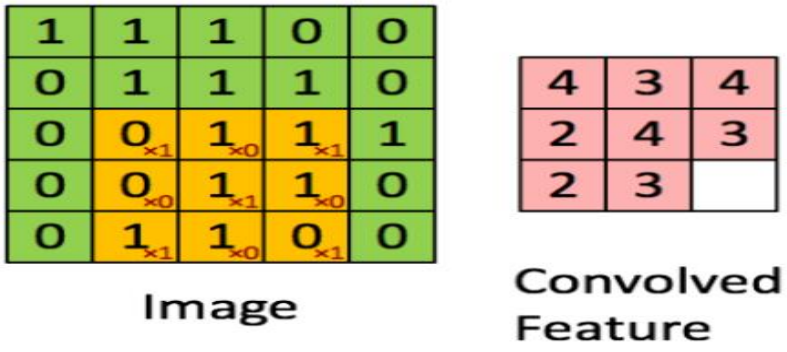
Giả sử chúng ta có dữ liệu đầu vào 2 chiều I, và ma trận nhân hai chiều K.

Phép tích chập có thể định nghĩa là:

$$S(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I(i+m, j+n)K(m,n)$$

(2.41)

Để dễ hình dung, ta có thể xem tích chập như một cửa sổ trượt (sliding window) áp đặt lên một ma trận. Cơ chế của tích chập qua hình minh họa:



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/#more-348

Hình 2.6: Phép tích chập

Ma trận bên trái là một bức ảnh đen trắng. Mỗi giá trị của ma trận tương đương với một điểm ảnh (pixel), 0 là màu đen, 1 là màu trắng (nếu là ảnh grayscale thì giá trị biến thiên từ 0 đến 255).

Cửa sổ trượt còn gọi tên là nhân, bộ lọc. Ở đây, ta dùng một ma trận bộ lọc 3x3 nhân từng thành phần tương ứng (element-wise) với ma trận bên trái. Giá trị đầu ra do tích của các thành phần này cộng lại. Kết quả của tích chập là một ma trận sinh ra từ việc trượt ma trận bộ lọc và thực hiện tích chập cùng lúc lên toàn bộ ma trận ảnh bên trái.

CNNs chỉ đơn giản bao gồm một vài tầng convolution kết hợp với các hàm kích hoạt phi tuyến (nonlinear activation function) như ReLU hay tanh để tạo ra thông tin trừu tượng hơn cho các tầng tiếp theo.

Trong mô hình mạng nơ-ron truyền thẳng (FNN), các tầng kết nối trực tiếp với nhau thông qua một trọng số w. Các tầng này còn được gọi là kết nối đầy đủ (full connected layer).

Trong mô hình CNNs thì ngược lại. Các tầng liên kết được với nhau thông qua cơ chế tích chập. Tầng tiếp theo là kết quả tích chập từ tầng trước đó, nhờ vậy mà ta có được các kết nối cục bộ. Nghĩa là mỗi nơ-ron ở tầng tiếp theo sinh ra từ bộ lọc áp đặt lên một vùng ảnh cục bộ của nơ-ron tầng trước đó.

Mỗi tầng như vậy được áp đặt các bộ lọc khác nhau, thông thường có vài trăm đến vài nghìn bộ lọc như vậy. Một số tầng khác như tầng pooling/subsampling dùng để chặn lọc lại các thông tin hữu ích hơn (loại bỏ các thông tin nhiễu).

Trong suốt quá trình huấn luyện, CNNs sẽ tự động học được các thông số cho các bộ lọc. Ví dụ trong tác vụ phân lớp ảnh, CNNs sẽ cố gắng tìm ra các thông số tối ưu cho các bộ lọc tương ứng theo thứ tự raw pixel > edges > shapes > facial > higher-level features. Tầng cuối cùng dùng để phân lớp ảnh.

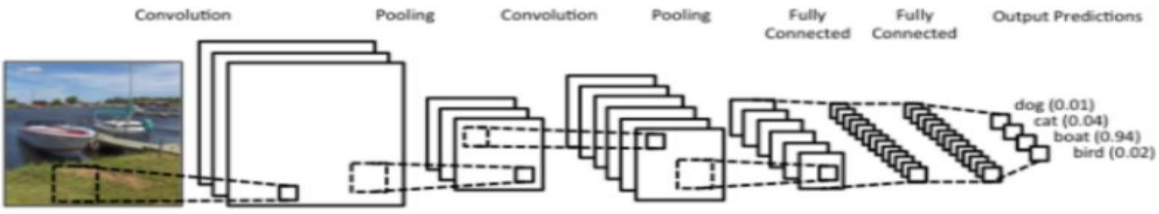


Image Classification with CNN

http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/#more-348

Hình 2.7: Mạng CNN

CNNs có tính bất biến và có tính kết hợp cục bộ (Location Invariance and Compositionality). Với cùng một đối tượng, nếu đối tượng này được chiếu theo các góc độ khác nhau (translation, rotation, scaling) thì độ chính xác của thuật toán sẽ bị ảnh hưởng đáng kể. Tầng Pooling sẽ cho bạn tính bất biến đối với phép dịch chuyển (translation), phép quay (rotation) và phép co giãn (scaling).

Tính kết hợp cục bộ cho ta các cấp độ biểu diễn thông tin từ mức độ thấp đến mức độ cao và trừu tượng hơn thông qua tích chập từ các bộ lọc. Đó là lý do tại sao CNNs cho ra mô hình với độ chính xác rất cao. Cũng giống như cách con người nhận biết các vật thể trong tự nhiên. Ta phân biệt được một con chó với một con mèo nhờ vào các đặc trưng từ mức độ thấp (có 4 chân, có đuôi) đến mức độ cao (dáng đi, hình thể, màu lông).

2.4.1. Tầng convolution

Xét l như là tầng convolution. Đầu vào tầng l bao gồm $m_1^{(l-1)}$ bản đồ đặc trưng từ tầng đằng trước, mỗi bản đồ có kích thước $m_2^{(l-1)}$ x $m_3^{(l-1)}$.

Tầng thứ i của bản đồ đặc trưng trong tầng l, xác định đầu ra Y_i^l:

$$Y_i^{(l)} = B_i^{(l)} + \sum_{j=1}^{m_1^{(l-1)}} K_{i,j}^{(l)} * Y_j^{(l-1)} \tag{2.42}$$

Trong đó:

$B_i^{(l)}$ là ma trận bias.

$K_{i,j}^{(l)}$ là kích thước bộ lọc $2h_1^{(l)}+1$ x $2h_2^{(l)}+1$ kết nối bản đồ đặc trưng thứ j trong tầng (l-1) với bản đồ đặc trưng thứ i trong tầng l.

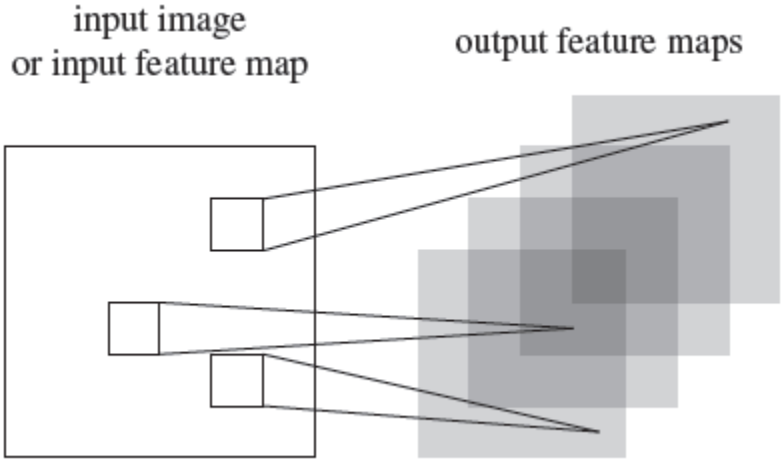
Khi đó đầu ra bản đồ đặc trưng tầng l:

$$m_2^{(l)} = m_2^{(l-1)}-2h_1^{(l)} \text{ và } m_3^{(l)} = m_3^{(l-1)}-2h_2^{(l)}$$

Thông thường các bộ lọc để tính toán một bản đồ đặc trưng $Y_i^{(l)}$ là giống nhau. Điều đó có nghĩa là: $K_{i,j}^{(l)} = K_{i,k}^{(l)}$ với mọi $j \neq k$.

Mỗi vị trí (r,s) trong ma trận $Y_i^{(l)}$ được tính bằng công thức:

$$(Y_i^{(l)})_{r,s} = (B_i^{(l)})_{r,s} + \sum_{j=1}^{m_1^{(l-1)}} (K_{i,j}^{(l)} * Y_j^{(l-1)})_{r,s} = (B_i^{(l)})_{r,s} + \sum_{j=1}^{m_1^{(l-1)}} \sum_{u=-h_1^{(l)}}^{h_1^{(l)}} \sum_{v=-h_2^{(l)}}^{h_2^{(l)}} (K_{i,j}^{(l)})_{u.v} * (Y_j^{(l-1)})_{r+u,s+v} \quad (2.43)$$



https://davidstutz.de/wordpress/wp-content/uploads/2014/07/seminar.pdf

Hình 2.8: Minh họa một tầng đơn convolution.

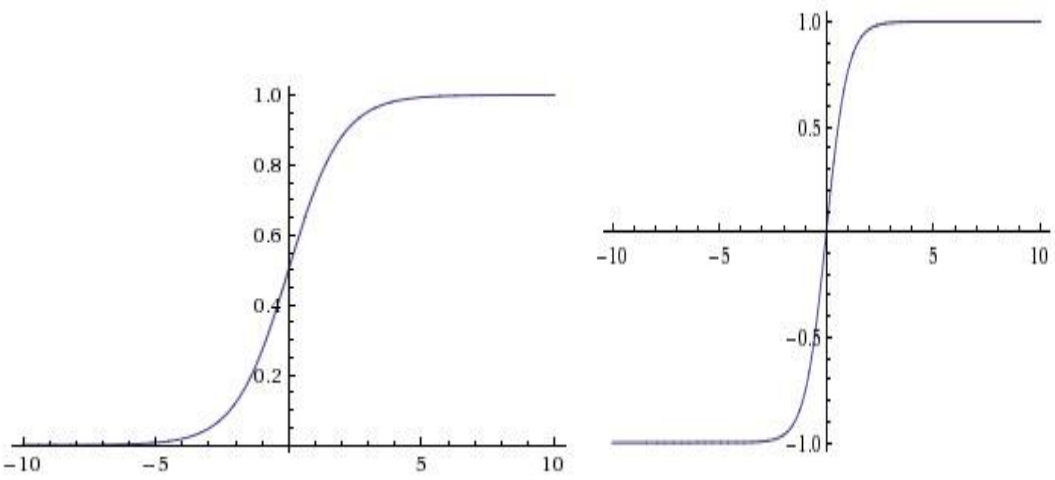
Tầng đầu vào hay là bản đồ trọng số của tầng đằng trước được nhân chập bởi các bộ lọc khác nhau để đưa ra một bản đồ đặc trưng của tầng l.

2.4.2. Tầng phi tuyến

Nếu l là tầng phi tuyến, đầu vào là $m_1^{(l)}$ bản đồ đặc trưng, đầu ra lặp lại $m_1^{(l)} = m_1^{(l-1)}$ bản đồ đặc trưng. Mỗi bản đồ kích thước $m_2^{(l)}$ x $m_3^{(l)}$, và giá trị tính bằng công thức:

$$Y_i^{(l)} = f(Y_i^{(l-1)})$$

Trong đó f là hàm phi tuyến như hàm sigmoid hay tanh.



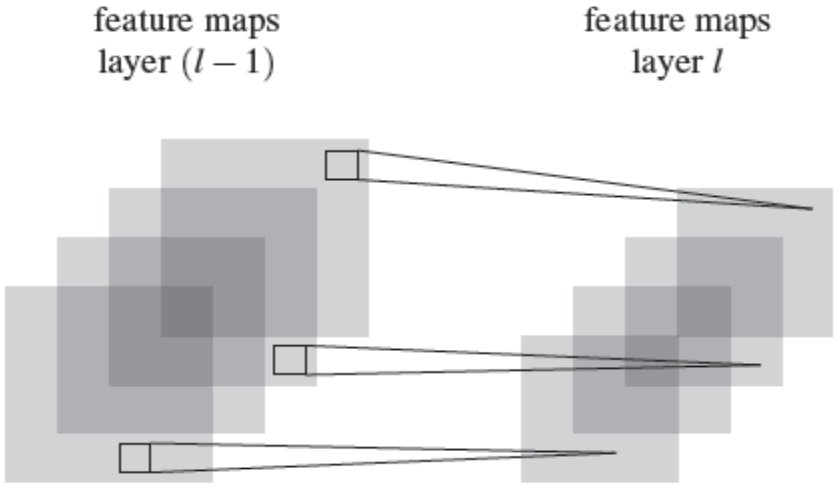
Hình 2.9: Hàm sigmoid (trái) Hàm tanh (phải)

2.4.3. Tầng pooling

Coi l là tầng pooling. Đầu ra được hợp thành từ $m_1^{(l)} = m_1^{(l-1)}$ bản đồ đặc trưng đã giảm kích thước. Tầng pooling thay thế các cửa sổ trượt tại các vị trí không chồng chéo trong mỗi bản đồ đặc trưng và giữ mỗi một giá trị cho mỗi cửa sổ như là việc bản đồ đặc trưng được lấy mẫu. Có hai kiểu pooling:

Average pooling: Lấy giá trị trung bình mỗi cửa sổ được chọn.

Max pooling: Lấy giá trị lớn nhất mỗi cửa sổ được chọn.



https://davidstutz.de/wordpress/wp-content/uploads/2014/07/seminar.pdf

Hình 2.10: Minh họa tầng pooling.

Coi l là tầng pooling và chọn $m_1^{(l-1)} = 4$ bản đồ đặc trưng của tầng trước. Tất cả các bản đồ đặc trưng được pooling và lấy mẫu độc lập. Mỗi đầu ra trong số $m_1^{(l)}$ bản đồ đặc trưng đưa ra một giá trị trung bình hoặc giá trị lớn nhất trong một cửa sổ cố định tương ứng với bản đồ đặc trưng trong tầng (l-1).

2.4.4. Tầng kết nối đầy đủ

Coi l là tầng kết nối đầy đủ. l lấy $m_1^{(l-1)}$ bản đồ đặc trưng kích thước $m_2^{(l-1)}$ x $m_3^{(l-1)}$ như đầu vào. Vị trí thứ i trong tầng l được tính bằng công thức:

$$y_i^{(l)} = f(z_i^{(l)})$$

Trong đó:

$$z_i^{(l)} = \sum_{j=1}^{m_1^{(l-1)}} \sum_{r=1}^{m_2^{(l-1)}} \sum_{s=1}^{m_3^{(l-1)}} w_{i,j,r,s}^{l} (Y_j^{(l-1)})_{r,s} \quad (2.44)$$

Với $w_{i,j,r,s}^{l}$ xác định trọng số kết nối giữa vị trí (r,s) tại bản đồ đặc trưng thứ j của tầng (l-1) và thứ i của tầng l.

Trong thực thế, tầng convolution được sử dụng để học các đặc trưng kế thừa và một hay nhiều tầng kết nối đầy đủ sử dụng cho mục đích phân loại dựa vào tính toán đặc trưng.

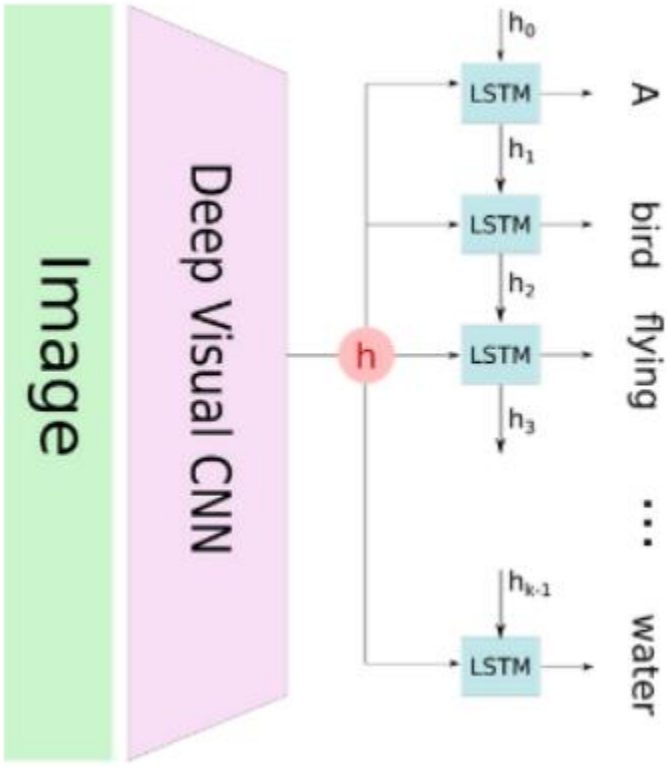
Lưu ý rằng, một tầng kết nối đầy đủ đã bao gồm hàm phi tuyến trong khi ở tầng convolution, tầng phi tuyến được tách rời trong lớp riêng của chúng.

Chương 3: Mô hình đề xuất

Các quá trình xử lý mạng nơ-ron liên quan đến Attention đã được nghiên cứu nhiều trong lĩnh vực thần kinh học. Các nghiên cứu liên quan là hiện thực hóa Attention: rất nhiều loại động vật tập trung trong việc xác định thành phần cụ thể đầu vào để tính toán phản hồi phù hợp. Nguồn gốc có một lượng lớn ảnh hưởng đến khoa học thần kinh khi chúng ta phải lựa chọn những thông tin phù hợp nhất, hơn là việc sử dụng tất cả các thôn tin, chứa một lượng lớn các thông tin không phù hợp cho phản hồi nơ-ron. Ý tưởng tập trung vào các thành phần cụ thể của đầu vào được áp dụng trong các ứng dụng của học sâu như nhận dạng tiếng nói, dịch máy, lý giải và nhận dạng thị giác của đối tượng.

Bài toán mở đầu là: Sinh một tiêu đề cho ảnh.

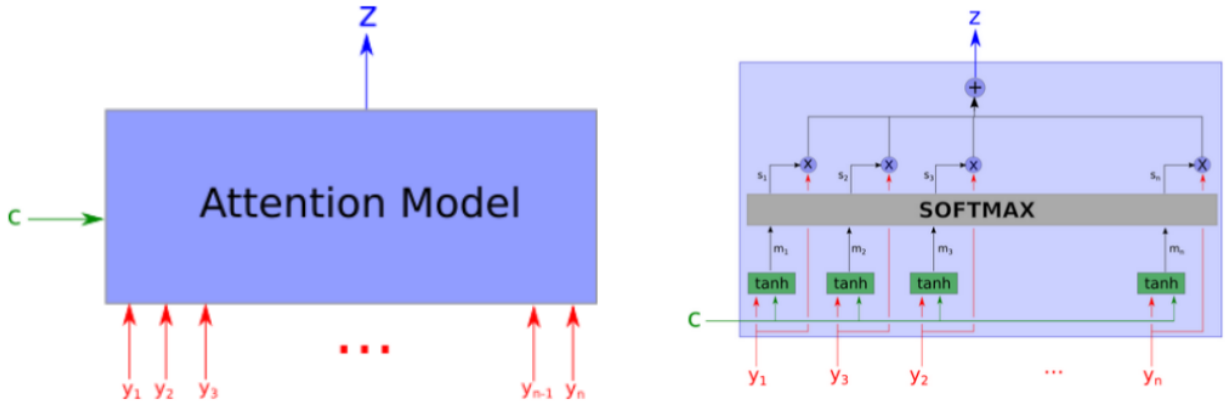
Một hệ thống cổ điển sinh tiêu đề có thể mã hóa hình ảnh, sử dụng một quá trình tiền xử lý CNN có thể đưa ra tầng ẩn h. Sau đó, nó có thể giải mã tầng ẩn bằng một mạng RNN, và sinh ra một đệ quy mỗi từ của tiêu đề.



https://blog.heuritech.com/2016/01/20/attention-mechanism/
Hình 3.1: Bài toán sinh tiêu đề

Vấn đề đặt ra với phương pháp này là khi mô hình cố gắng sinh ra một từ của tiêu đề, từ này thường chỉ mô tả duy nhất một phần của hình ảnh. Sử dụng toàn bộ biểu diễn của bức ảnh h để điều kiện hóa việc sinh mỗi từ sẽ không hiệu quả trong việc đưa ra một từ khác cho những thành phần khác của bức ảnh. Điều này lý giải cho lợi ích của kỹ thuật Attention.

Với kỹ thuật Attention, bức ảnh đầu tiên được chia thành n thành phần và chúng có thể tính toán với sự trình diễn CNN cho mỗi thành phần $h_1,...,h_n$. Khi RNN sinh ra một từ mới, kỹ thuật Attention tập trung vào những thành phần phù hợp của bức ảnh, vì thế quá trình giải mã chỉ sử dụng thành phần cụ thể của ảnh.



https://blog.heuritech.com/2016/01/20/attention-mechanism/
Hình 3.2: Sơ đồ mô hình Attention

Trước khi sử dụng Cơ chế Attention, các mô hình tóm tắt đều có cơ chế sử dụng Encoder-Decoder. Tại bước encoder, đầu vào của mạng RNN, LSTM, GRU là các vector được tạo ra từ mã hóa chuỗi từ với mô hình từ nhúng (word embedding). Pha decoder sử dụng một mạng RNN, LSTM hoặc GRU tương ứng để sinh ra một chuỗi từ mới dựa vào chuỗi đầu vào và các từ sinh ra phía trước. Trong mô hình tóm tắt văn bản tự động, thay vì tìm ra xác suất lớn nhất của mỗi từ sinh ra ở bước decoder, chúng ta tạo ra danh sách các từ ứng viên tại mỗi bước giải mã. Sau đó sử dụng giải thuật tìm kiếm chùm (Beam Search) để lựa chọn các từ ứng viên và kết nối danh sách các từ ứng viên đó lại thành một câu có điểm số cao nhất tạo ra một chuỗi tóm tắt.

3.1. Cơ chế Attention

3.1.1. Kiến trúc RNN Encoder-Decoder

Được đề xuất bởi Cho[12] và Sutskever[10] như là một kiến trúc hiện đại có thể học sự căn chỉnh và dịch ngay lập tức.

Trong Encoder-Decoder, một encoder đọc vào một câu - một chuỗi vector x = (x_1,…,x_{Tx}) thành một vector c. Cách tiếp cận như sau:

$$h_t = f(x_t, h_{t-1}) \tag{3.1}$$

$$c = q(\{h_1,\ldots,h_{Tx}\}) \tag{3.2}$$

Trong đó h_t là trạng thái ẩn tại thời điểm t, $h_t \in \mathbb{R}^n$ và c là vector được sinh ra từ một chuỗi các trạng thái ẩn. f và q là các hàm phi tuyến.

Pha decoder, được huấn luyện để dự đoán từ tiếp theo $y_{t'}$ cho ngữ cảnh c và tất cả các từ dự đoán đằng trước $\{y_1,\ldots,y_{t'-1}\}$. Hiểu theo cách khác decoder định nghĩa một xác xuất trên chuyển dịch y bằng việc phân tích xác suất liên kết thành thứ tự các điều kiện:

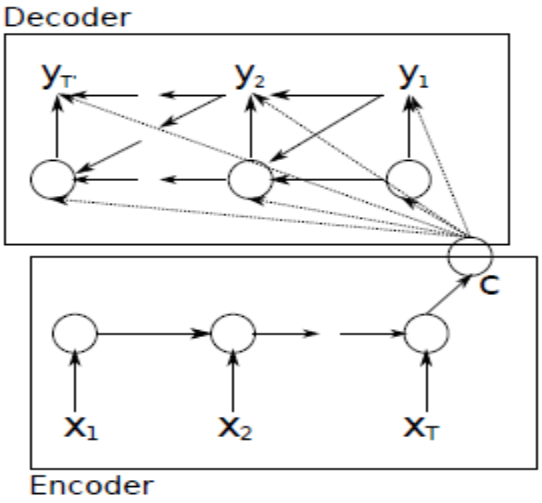
$$p(y) = \prod_{t=1}^{T} p(y_t|\{y_1, \ldots, y_{t-1}\}, c) \tag{3.3}$$

Trong đó y = (y_1,…y_{Ty}).

Với một mạng RNN, mỗi xác suất có điều kiện được mô hình bởi:

$$p(y_t|y_1, \ldots, y_{t-1}, c) = g(y_{t-1}, s_t, c) \tag{3.4}$$

Trong đó g là hàm phi tuyến, y_t là đầu ra và s_t là trạng thái ẩn của mạng RNN.



Kyunghyun Cho et al. [12]

Hình 3.3: Minh họa kiến trúc của mạng Encoder-Decoder

3.1.2. Cơ chế Attention

Kiến trúc Encoder-Decoder có thể bị phá vỡ khi chuỗi đầu vào quá dài. Nguyên nhân là nếu ở mỗi bước nếu chỉ có một vector ngữ cảnh c giao tiếp giữa encoder và decoder, vector đó sẽ phải mã hóa cho toàn bộ chuỗi đầu vào, dẫn đến nó có thể bị tan biến khi nó xử lý chuỗi ký tự quá dài. Cơ chế Attention cho phép bộ giải mã tập trung vào một phần khác nhau từ đầu ra của encoder.

Định nghĩa mỗi xác suất có điều kiện như sau:

$$p(y_i|\{y_1, \dots, y_{i-1}\}, x) = g(y_{i-1}, s_i, c_i) \tag{3.5}$$

Trong đó:

Mỗi s_i là một trạng thái ẩn RNN tại thời điểm i, tính bằng công thức:

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \tag{3.6}$$

Điều này không giống với cách tiếp cận encoder-decoder, ở đây mỗi xác suất được điều kiện trên một ngữ cảnh riêng biệt c_i cho mỗi từ mục tiêu y_i.

Vector ngữ cảnh c_i phụ thuộc vào chuỗi trạng thái (h_1,…,h_{Tx}) – để encoder ánh xạ câu đầu vào. Mỗi trạng thái h_i chứa đựng thông tin của toàn bộ câu với một sự nhấn mạnh các thành phần xung quanh từ thứ i của câu đầu vào.

Ngữ cảnh c được tính toán như là trọng số tổng hợp của các trạng thái h_i:

$$c_i = \sum_{j=1}^{T_x} \alpha_{i,j} h_j \tag{3.7}$$

Trong đó: trọng số $\alpha_{i,j}$ của mỗi trạng thái h_j được tính như sau:

$$\alpha_{i,j} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{3.8}$$

Với $e_{ij} = a(s_{i-1}, h_j)$ là hình thức căn lề tính điểm khả năng đầu vào xung quanh vị trí j và đầu ra tại vị trí i trùng nhau. Điểm số dựa trên trạng thái ẩn RNN s_{i-1} và trạng thái gán nhãn h_j của câu đầu vào.

Xác suất α_{ij} hay e_{ij} phản ánh độ quan trọng của trạng thái h_j với trạng thái ẩn đằng trước s_{i-1} để quyết định trạng thái tiếp theo s_i và đưa ra nhãn y_i. Decoder quyết định thành phần của câu đầu vào để tập trung. Encoder toàn bộ thông tin câu thành một vector có độ dài cố định. Thông tin có thể trải dài thành chuỗi gán nhãn, có thể lựa chọn lấy lại bởi pha decoder tương ứng.

Toàn bộ mô hình được huấn luyện end-to-end bằng việc cực tiểu hóa xác suất có điều kiện:

$$L(\theta) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_y^n} logp(y_t = y_t^n | y_{<t}^n, X^n) \tag{3.9}$$

Trong đó: N là số lượng các cặp câu, X^n là câu đầu vào, y_t^n là nhãn đầu ra thứ t trong n cặp tương ứng.

3.1.3. BiRNN

Đối với rất nhiều nhiệm vụ gán nhãn chuỗi, việc truy cập vào thông tin tương lai rất có ích cho bối cảnh quá khứ. Ví dụ, khi phân loại một chữ viết tay, sẽ rất hữu ích khi biết chữ cái đến từ đằng sau cũng như chữ cái đến từ đằng trước nó. Tuy vậy, mạng RNN chuẩn xử lý chuỗi theo thứ tự thời gian, chúng bỏ qua tương lai của ngữ cảnh. Một giải pháp rõ ràng là thêm một cửa sổ trượt của ngữ cảnh tương lai vào mạng đầu vào. Tuy nhiên, nó làm tăng số lượng bộ trọng số đầu vào. Một cách tiếp cận khác là tạo sự trễ giữa các yếu tố đầu vào và mục tiêu, nhờ đó tạo cho mạng một số mốc thời gian của ngữ cảnh tương lai. Phương pháp này tuy vẫn duy trì được điểm mạnh của mạng RNN đối với sự biến dạng, nhưng nó vẫn yêu cầu phạm vi của ngữ cảnh phải xác định bằng tay. Hơn thế nữa nó đặt một gánh nặng không cần thiết lên mạng bằng cách buộc nó phải nhớ bản gốc đầu vào và bối cảnh trước đó của nó, trong suốt thời gian trễ. Trong các phương án trên, không có phương pháp nào loại bỏ sự không cân xứng giữa thông tin quá khứ và tương lai.

Mạng hai chiều RNN (BiRNN) được đưa ra như một giải pháp phù hợp. Ý tưởng cơ bản của BiRNN là trình bày mỗi chuỗi tiến và chuỗi lùi thành hai tầng ẩn hồi quy riêng biệt, cả hai đều được kết nối với nhau tới một tầng giống nhau. Cấu trúc này cung cấp cho mỗi tầng đầu ra với quá khứ hoàn chỉnh và bối cảnh tương lai cho mọi điểm trong chuỗi đầu vào, mà không phải di dời các đầu vào từ các mục tiêu phù hợp. BiRNN đã cải thiện kết quả trong các lĩnh vực khác nhau, chúng hoạt động tốt hơn RNN một chiều khi gán nhãn chuỗi.

Thông thường RNN đọc câu đầu vào theo thứ tự bắt đầu của câu từ điểm đầu tiên x_1 tới điểm cuối x_{Tx}. BiRNN được đề xuất để tổng hợp mỗi từ không chỉ đằng trước một từ mà còn từ đằng sau từ đó.

BiRNN bao gồm chiều tiến RNN và chiều quay lui RNN. Chiều tiến $\vec{h}$ RNN đọc câu đầu vào theo đúng thứ tự (từ x_1 đến x_{Tx}) và tính toán trạng thái ẩn ($\overrightarrow{h_1},\ldots,\overrightarrow{h_{T_x}}$) Chiều quay lui RNN $\overleftarrow{h}$ đọc câu đầu vào theo thứ tự ngược lại (từ x_{Tx} tới x_1). Kết quả trong chuỗi quay lui trạng thái ẩn ($\overleftarrow{h_1},\ldots,\overleftarrow{h_{T_x}}$).

Để đạt được trạng thái cho mỗi từ x_j, ta kết nối chiều trạng thái tiến $\vec{h}$ và chiều quay lui $\overleftarrow{h}$.

$$h_j = [\overrightarrow{h_j^T}; \overleftarrow{h_j^T}]$$

(3.10)

Trạng thái gán nhãn h_j bao gồm thông tin tổng hợp của cả đằng trước và đằng sau từ đó. Phụ thuộc vào xu hướng RNN trình bày câu gần từ mà trạng thái ẩn h_j sẽ tập trung xung quanh từ x_j. Chuỗi trạng thái được sử dụng bởi decoder và model căn chỉnh để tính toán vector ngữ cảnh.

Pha tiến của tầng ẩn BiRNN giống như mạng RNN chuẩn, trừ việc chuỗi đầu ra được trình bày theo các hướng ngược nhau với hai lớp ẩn, tầng đầu ra không được cập nhật cho đến khi cả hai tầng ẩn đã được xử lý toàn bộ chuỗi đầu vào.

for $t = 1$ to T **do**
  Forward pass for the forward hidden layer, storing activations at each timestep
for $t = T$ to 1 **do**
  Forward pass for the backward hidden layer, storing activations at each timestep
for all t, in any order **do**
  Forward pass for the output layer, using the stored activations from both hidden layers

Alex Graves [21]

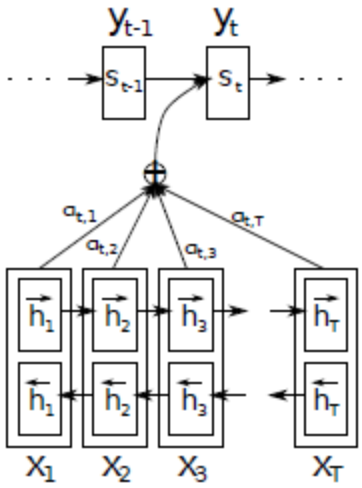
Hình 3.4: Pha tiến của mạng BiRNN

Tương tự quá trình quay lui như với một mạng RNN chuẩn trừ việc tất cả các tầng đầu ra δ được tính toán đầu tiên và sau đó quay trở lại hai tầng ẩn theo hướng ngược lại.

for all t, in any order **do**
  Backward pass for the output layer, storing δ terms at each timestep
for $t = T$ to 1 **do**
  BPTT backward pass for the forward hidden layer, using the stored δ terms from the output layer
for $t = 1$ to T **do**
  BPTT backward pass for the backward hidden layer, using the stored δ terms from the output layer

Alex Graves [21]

Hình 3.5: Pha quay lui của mạng BiRNN



Dzmitry Bahdanau et *al. [9]*.

Hình 3.6: Minh họa cơ chế Attention

3.2. Thuật toán tìm kiếm chùm (Beam search)

Trong mô hình tóm tắt, bộ giải mã được điều khiển bởi một câu đã được mã hóa để tạo ra câu mới. Tại mỗi bước lặp t, bộ giải mã cần đưa ra quyết định từ nào sinh ra từ thứ t trong câu. Vấn đề là chúng ta không biết chính xác chuỗi từ cần sinh ra để cực đại hóa xác suất có điều kiện tổng thể. Để giải quyết vấn đề này thuật tìm kiếm chùm sẽ được áp dụng. Thuật toán có độ rộng K sao cho tại mỗi bước đưa ra K đề xuất và tiếp tục giải mã với một trong số chúng.

Các mô hình phát triển giải quyết vấn đề sinh chuỗi thường hoạt động bằng sinh ra các phân phối xác suất thông qua từ điển các từ đầu ra. Chúng ta đối mặt với vấn đề này lúc làm việc với mạng nơ-ron truy hồi (RNN), khi mà văn bản được sinh ra như đầu ra. Ở tầng cuối cùng trong mạng nơ-ron có một mạng nơ-ron cho mỗi từ trong từ điển đầu ra và một hàm kích hoạt được sử dụng để đưa ra khả năng mỗi từ trong từ vựng là từ tiếp theo trong chuỗi.

Pha giải mã liên quan đến tìm kiếm thông qua tất cả các chuỗi đầu ra dựa trên khả năng của chúng. Kích thước tập từ vựng có thể tới hàng ngàn, hàng triệu từ. Vì thế vấn đề tìm kiếm là số mũ trong chiều dài cả chuỗi đầu ra và là vấn đề NP khó để hoàn tất tìm kiếm.

Thông thường, các phương pháp tìm kiếm thông minh được sử dụng để đưa ra chuỗi đầu ra được giải mã gần đúng cho sự dự đoán. Chuỗi ứng viên của các từ được ghi điểm dựa trên khả năng của chúng. Phương pháp phổ biến là tìm kiếm tham lam hoặc tìm kiếm chùm để định vị chuỗi ứng viên của văn bản.

Khác với các phương pháp thông minh, thuật toán tìm kiếm chùm mở rộng trên thuật toán tham lam và trả về danh sách phù hợp nhất các chuỗi đầu ra. thay vì tham lam chọn bước tiếp theo có khả năng nhất khi chuỗi được xây dựng, thuật toán tìm kiếm chùm mở rộng các khả năng có thể ở bước kế tiếp và giữa k trường hợp phù hợp nhất, trong đó k là tham số người dùng chỉ định và kiểm soát số lượng các chùm hoặc tìm kiếm song song thông qua chuỗi xác suất.

Thông thường độ rộng chùm là 1 tương ứng với thuật toán tìm kiếm tham lam và giá trị 5 hoặc 10 cho tiêu chuẩn chung của dịch máy. Độ rộng chùm kết quả lớn hơn dẫn tới hiệu suất tốt hơn của một mô hình vì các chuỗi ứng viên nhiều khả năng làm tăng khả năng kết hợp tốt hơn một chuỗi mục tiêu. Sự tăng hiệu suất này làm giảm tốc độ giải mã.

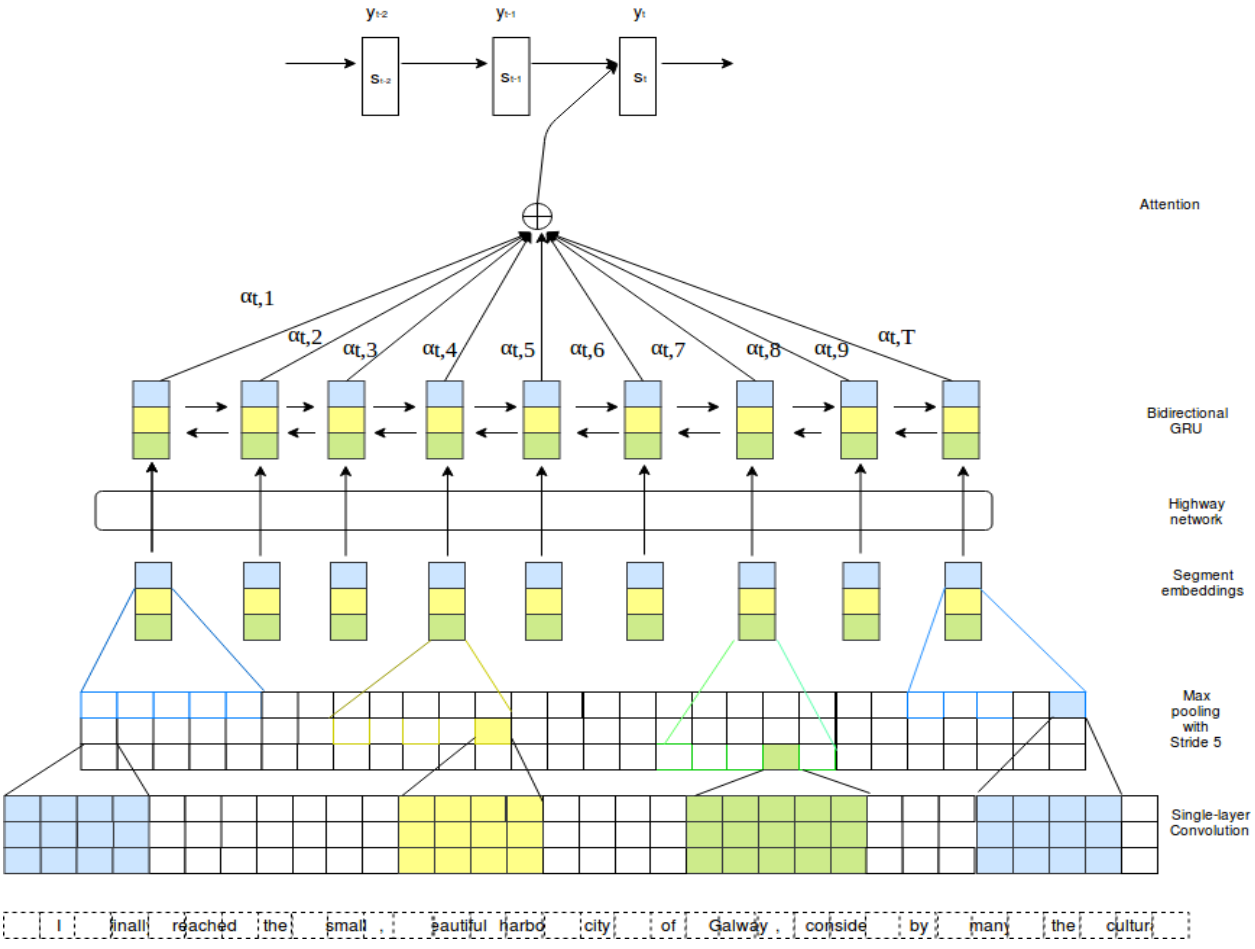
Cho (2014)[15] đã cài đặt một thuật toán tìm kiếm chùm tiêu chuẩn trong pha giải mã của dịch máy (Koehn, 2004) cho một hệ thống Encoder-Decoder trong GroundHog. Thuật toán chùm đã thành công trong việc giảm thiểu không gian tìm kiếm từ kích thước mũ sang kích thước đa thức.

Cho một pha encoder, một pha decoder và một đầu vào là x, chúng ta tìm kiếm chuỗi dịch tốt nhất $\hat{y} = \text{argmax}_y p(y|x)$. Một nhóm các ngăn xếp được sử dụng để lưu lại các giả thuyết trong quá trình tìm kiếm. Kích thước chùm N được sử dụng để điều kiển không gian tìm kiếm bằng việc mở rộng đỉnh N giả thuyết trong ngăn xếp hiện tại. Với những cài đặt bên trên, phần dịch y được sinh ra từ bởi từ theo chiều từ trái sang phải. Ta định nghĩa một giả thuyết hoàn tất là câu chứa đầu ra EOS, trong đó EOS là từ đặc biệt chỉ ra kết thúc trong câu.

3.3. Mô hình đề xuất

Các mô hình học sâu áp dụng trong bài toán tóm tắt văn bản gồm: nhóm tác giả Rush [2] sử dụng mạng nơ-ron tích chập kết hợp với cơ chế attention. Sau đó, nhóm Chopra [3] sử dụng mạng nơ-ron tích chập và mạng RNN kết hợp với cơ chế attention. Nhóm Nallapati[19] sử dụng mô hình GRU và cơ chế attention đạt kết quả tốt hơn nhóm của Rush [2]. Hơn nữa mô hình của nhóm tác giả Nallapati[19] còn áp dụng được trên cả dữ liệu tóm tắt chứa nhiều câu văn. Điều mà nhóm tác giả Rush[2] và Chopra[3] chưa tiến hành thí nghiệm.

Do đó, tôi mở rộng nghiên cứu của nhóm tác giả Nallapati[19] bằng cách sử dụng mạng nơ-ron tích chập với mạng GRU kết hợp với cơ chế Attention. Câu đầu vào được đi qua các tầng Convolution rồi đến tầng mạng Highway. Đầu ra của tầng Hìghway sẽ là đầu vào của mạng GRU và đi vào cơ chế Attention.



Hình 3.7: Mô hình đề xuất

Tầng nhúng (embedding): Giả sử ta có câu nguồn $X = (x_1, x_2, ... x_{Tx}) \in R^{d \times Tx}$.

Trong đó: d là số chiều của một từ.

Tầng convolution:

Giả định ta có một hàm lọc $f \in R^{d \times w}$ với độ rộng là w, đầu tiên chúng ta áp dụng biên ở đầu và cuối của câu X. Do đó, biên của câu tạo thành $X' \in R^{d \times (Tx + w - 1)}$ là w-1 từ. Ta áp dụng phép tích chập giữa X' và f sao cho phần tử đầu ra thứ k được tính như sau:

$$Y_k = (X' * f)_k = \sum_{i,j} (X'_{[:,k-w+1:k]} \otimes f)_{ij} \quad (3.11)$$

Trong đó:

$\otimes$ là phép nhân từng phần ma trận và phép toán * là phép tích chập. $X'_{[:,k-w+1:k]}$ là một tập con của X' chứa tất cả các hàng nhưng chỉ chứa w cột kề bên. Kiểu lựa chọn lề như vậy gọi là một nửa tích chập (half convolution). Điều này đảm bảo chiều dài của đầu ra là $Y \in R^{1xTx}$.

Bên trên, ta minh họa trường hợp một bộ lọc tích chập cố định. Để trích chọn các mẫu thông tin với chiều dài khác nhau, ta đưa một tập các bộ lọc với chiều dài khác nhau. Cụ thể hơn, ta sử dụng một tập các bộ lọc $F = \{f_1, ..., f_m\}$. Trong đó, $f_i = R^{d \times i \times ni}$ là một tập của các n_i bộ lọc với độ rộng i. Mô hình của tôi sử dụng m=5, do đó có thể trích chọn được 5 gram chiều dài. Đầu ra của tất cả các hàm lọc được xếp chồng lại, đưa ra một sự biểu diễn đơn giản $Y \in R^{NxTx}$, trong đó số chiều của mỗi cột được cho bởi tổng các bộ lọc $N = \sum_{i=1}^{m} n_i$. Cuối cùng tầng kích hoạt được áp dụng theo từng phần tử của sự trình diễn.

Tầng max pooling:

Đầu ra của tầng convolution đầu tiên được phân thành các cụm với chiều dài là s, và tầng max pooling được áp dụng với mỗi cụm không giao nhau. Thủ tục lựa chọn các đặc trưng nổi bật nhất đưa ra một phân đoạn nhúng. Mỗi tầng nhúng là

một tóm tắt của một đoạn riêng biệt (hoặc chồng chéo) trong câu đầu vào. Điều này hoạt động như đơn vị ngôn ngữ bên trong từ tầng hiện tại đến tầng trên.

Sự rút ngắn sự biểu diễn nguồn theo s-fold: $Y' \epsilon R^{Nx(Tx/s)}$. Theo kinh nghiệm, tôi sử dụng s=5.

Mạng highway (nhóm tác giả Srivastava 2015 [14]):

Mạng highway được áp dụng khi số tầng của mô hình học sâu tăng lên cùng với đó là sự tăng độ phức tạp tính toán. Mạng highway có thể sử dụng với hàng trăm tầng được huấn luyện trực tiếp cùng với phương pháp tối ưu SGD và các biến thể của hàm kích hoạt.

Chuỗi ma trận nhúng sau khi qua tầng max pooling của mạng nơ-ron tích chập được đưa đến mạng highway. Ý nghĩa tiềm ẩn là mạng highway chuyển đổi đầu ra của tầng max pooling thành các khoảng ngữ nghĩa, giúp các đặc trưng được học chính xác. Mạng này chuyển đổi đầu vào x với một cơ chế cổng để điều chỉnh thông tin theo luồng:

$$y = g \odot ReLU(Wx + b) + (1 - g) \odot x \tag{3.12}$$

Đầu ra của tầng mạng highway được đưa tới mạng GRU hai chiều.

Cuối cùng, một tầng mạng hướng tiến tính toán điểm số attention của mỗi từ mục tiêu để sản sinh cho mỗi cụm thể hiện đầu vào.

Chương 4: Thực nghiệm và đánh giá

4.1. Dữ liệu thử nghiệm

Tôi sử dụng hai bộ dữ liệu để tiến hành thí nghiệm: Bộ dữ liệu Gigaword và bộ dữ liệu CNN/Daily Mail.

4.1.1. Bộ dữ liệu Gigaword

Bộ dữ liệu đầu tiên lấy tại địa chỉ: https://github.com/harvardnlp/sent-summary.

Dữ liệu này bao gồm dữ liệu Gigaword chứa khoảng 3.8 triệu cặp câu gồm câu nguồn và câu tóm tắt từ dữ liệu CNN và Dailymail. Chúng cũng chứa dữ liệu DUC 2003 và DUC 2004.

Bảng 4.1. Thống kê dữ liệu Gigaword

	Gigaword			DUC2003	DUC2004
	Huấn luyện	Phát triển	Kiểm thử		
Số lượng câu	38039957	189651	1951	624	500

Tập kiểm thử Gigaword chứa 1 file dữ liệu gốc và 1 file do con người đánh giá.

Tập kiểm thử của DUC2003 và DUC2004 chứa 1 file dữ liệu gốc và 3 file do người dùng đánh giá tương ứng.

Bảng 4.2. Ví dụ dữ liệu Gigaword

Câu nguồn	australia 's current account deficit shrunk by a record #.## billion dollars -lrb- #.## billion us -rrb- in the june quarter due to soaring commodity prices , figures released monday showed .
Câu tóm tắt	australian current account deficit narrows sharply
Câu nguồn	at least two people were killed in a suspected bomb attack on a passenger bus in the strife-torn southern philippines on monday , the military said .

Câu tóm tắt	at least two dead in southern philippines blast
Câu nguồn	australian shares closed down #.# percent monday following a weak lead from the united states and lower commodity prices , dealers said .
Câu tóm tắt	australian stocks close down #.# percent
Câu nguồn	south korea 's nuclear envoy kim sook urged north korea monday to restart work to disable its nuclear plants and stop its `` typical " brinkmanship in negotiations .
Câu tóm tắt	envoy urges north korea to restart nuclear disablement
Câu nguồn	south korea on monday announced sweeping tax reforms , including income and corporate tax cuts to boost growth by stimulating sluggish private consumption and business investment .
Câu tóm tắt	skorea announces tax cuts to stimulate economy

4.1.2. Bộ dữ liệu CNN/Daily Mail

Bộ dữ liệu thứ hai, tôi sử dụng dữ liệu huấn luyện của nhóm tác giả Jianpeng Cheng[20].

Dữ liệu gồm các bài báo trên CNN và Daily Mail. Mỗi nguồn bài báo chia thành 3 thư mục: Huấn luyện, phát triển và kiểm thử. Tôi gộp hai nguồn bài thành ba thư mục: Huấn luyện, phát triển và kiểm thử.

Bảng 4.3. Thống kê dữ liệu CNN/Daily Mail

	Huấn luyện	Phát triển	Kiểm thử
Dailymail	193986	12147	10350
CNN	83568	1220	1093
Tổng cộng	277554	13367	11443

Bảng 4.4. Ví dụ dữ liệu CNN/Daily Mail

Văn bản	CARACAS , Venezuela -- Venezuela president Chavez says he would be willing to accept prisoners from the Guantanamo detention center , which U.S. president Obama has said he will close , the Venezuela government said thursday president Obama has pledged to close the detention facility at Guantanamo , Cuba Chavez also said he hopes the U.S. will give Cuba back the land on which the naval base is located , the government said in a news release " we would not have any problem receiving a human being , " the government release quoted Chavez as saying in an interview wednesday with Al Jazeera the U.S. obtained the Guantanamo base in 1903 , after Spain 's surrender in the Spanish-American War of 1898 in 2002 , then - president Bush opened the detention center to hold what the Bush administration categorized as enemy combatants captured in Iraq , Afghanistan and elsewhere U.S. officials have not said what will happen to prisoners at the camp when it closes , nor are there are any known plans for any to be sent to Venezuela Chavez attended the second summit of South American and Arab heads of state in Qatar earlier this week speaking about Israel , Chavez said new prime minister Benjamin Netanyahu is supported " by the extreme right , " the government release said " i hope someday the Hebrew people will be liberated from that caste , " the release quoted him as saying in the 90 - minute Al Jazeera interview from Qatar , Chavez traveled to Iranian , where he met with president Mahmoud Ahmadinejad on thursday Maria Carolina Gonzalez contributed to this report for CNN .
Tóm tắt	Chavez would be willing to accept Guantanamo inmates , Venezuela says Venezuela president quoted as having no problem " receiving a human being " no plans are known for *sending* inmates to Venezuela when detention center closes on Middle East trip , Chavez *criticizes* Israel , meets with Iranian president
Văn bản	-- five people were killed and 10 critically injured saturday when a minivan crashed on I-10 near Baton Rouge , Louisiana , state police said fifteen people were in the minivan , said trooper Graham , and only two were wearing seat belts among the dead were children as young as 3 years old , he said " the minivan blew out a tire and the driver lost control , " Graham said the vehicle " sideswiped a box truck and then ran off the road into the left median , overturned multiple times and

	finally came to rest upright on the eastbound side of I-10 , " Graham said the one person in the truck was not injured the accident shut down I-10 in both directions shortly after 12:15 p.m. (1:15 p.m. et) ; one lane in each direction was opened about two hours later alcohol and drugs were not suspected factors in the crash , but blood was drawn from the driver -- one of the fatalities -- to confirm , Graham said the accident came soon after the Louisiana Legislature passed a law requiring riders in every seat to be buckled up " this is an example of why we implemented that law , " Graham said " it 's very frustrating for us to come out here and see children dead , " he said .
Tóm tắt	state police : 15 people were in the minivan , only two in *seatbelts* crash occurred after minivan blew a tire , trooper says crash near Baton Rouge shut down I-10 in both directions for about two hours

4.2. Cài đặt

Tôi sử dụng framework dl4mt cho bài toán dịch máy sử dụng cơ chế Attention với mạng GRU tại địa chỉ https://github.com/nyu-dl/dl4mt-tutorial.

Đối với bộ dữ liệu Gigaword, kích thước từ điển là 3000 từ. Số chiều của từ sử dụng là 300. Chiều dài câu tối đa là 100. Đối với bộ dữ liệu CNN/Daily Mail, kích thước từ điển đầu vào là 18000, kích thước từ điển đầu ra là 60000, số chiều của từ là 128, độ dài đoạn văn tối đa là 800.

Phương pháp tối ưu sử dụng là adadelta với hệ số học 0.0001. Tất cả bộ trọng số được khởi tạo trong phân phối chuẩn [-0.01, 0.01]. Pha decode, tôi sử dụng thuật toán beam search. Kích thước beam search là 20 cho tất cả mô hình.

Cấu hình server chạy: Ubuntu server, 32 core, 96G RAM. GPU Quadro K2200, bộ nhớ 4G.

Tôi sử dụng mạng CNN với các cấu hình như sau:

Bộ lọc 1: sử dụng 1 kiểu bộ lọc với số lượng bộ lọc mỗi kiểu là 200.

Bộ lọc 2: sử dụng 2 kiểu bộ lọc với số lượng bộ lọc mỗi kiểu là 200 – 250.

Bộ lọc 3: sử dụng 3 kiểu bộ lọc với số lượng bộ lọc mỗi kiểu là 200 – 250 – 300.

Bộ lọc 4: sử dụng 4 kiểu bộ lọc với số lượng bộ lọc mỗi kiểu là 200 – 250 – 300 – 300.

Bộ lọc 5: sử dụng 5 kiểu bộ lọc với số lượng bộ lọc mỗi kiểu là 200 – 250 – 300 – 300 – 400.

4.3. Kết quả

Để đánh giá kết quả của phương pháp, tôi sử dụng hệ thống độ đo ROUGE, được điều chỉnh bởi DUC như hệ thống ước lượng chính cho tóm tắt văn bản. Nó bao gồm năm độ đo, để xác định chất lượng bản tóm tắt bởi máy so với bản tóm tắt bởi con người, đó là: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S và ROUGE-SU. Sự đo lường thực hiện bởi số lượng đơn vị trùng lặp như N-grams, chuỗi các từ, cặp các từ giữa văn bản tóm tắt ứng cử và văn bản tóm tắt dẫn xuất.

ROUGE-N ước lượng độ phủ N-grams giữa văn bản tóm tắt ứng cử và văn bản tóm tắt dẫn xuất.

$$ROUGE - N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count_{match}(N - gram)}{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count(N - gram)} \quad (4.1)$$

Trong đó N là chiều dài của N-grams, $Count_{match}$(N-gram) là số lượng lớn nhất N-grams cùng xuất hiện giữa hai bản tóm tắt tương ứng, Count(N-gram) là số lượng N-grams trong trong văn bản tóm tắt dẫn xuất.

ROUGE-L sử dụng độ đo chuỗi con có độ dài lớn nhất (LCS – Longest Common Subsequence) để ước lượng sự tóm tắt. Mỗi câu được xem như chuỗi các từ và do đó LCS giữa văn bản tóm tắt ứng cử và văn bản tóm tắt dẫn xuất được xác định. ROUGE-L tính toán tỉ lệ giữa độ dài của LCS và chiều dài của văn bản tóm tắt dẫn xuất.

$$\begin{cases} P_{LCS}(R,S) = \dfrac{LCS(R,S)}{|S|} \\ R_{LCS}(R,S) = \dfrac{LCS(R,S)}{|R|} \\ R_{LCS}(R,S) = \dfrac{(1+\beta^2)P_{LCS}(R,S)R_{LCS}(R,S)}{\beta^2 P_{LCS}(R,S) + R_{LCS}(R,S)} \end{cases} \quad (4.2)$$

Trong đó:
|R| và|S| tương ứng là chiều dài văn bản dẫn xuất R và văn bản ứng viên S.
LCS(R,S) là LCS giữa R và S.
P_{LCS}(R,S) là độ chính xác của LCS(R,S) và R_{LCS}(R,S) là độ phủ của LCS(R,S).
β là P_{LCS}(R,S) / R_{LCS}(R,S).

4.3.1. Bộ dữ liệu Gigaword

Kết quả chạy với các cấu hình bộ lọc của mô hình CNN.

Bảng 4.5. Kết quả với dữ liệu Gigaword

	RG-1	RG-2	RG-L
Bộ lọc 1	25.86	8.69	23.95
Bộ lọc 2	25.54	8.78	23.78
Bộ lọc 3	**27.00**	**9.62**	**24.70**
Bộ lọc 4	26.62	9.23	24.49
Bộ lọc 5	26.75	9.47	24.79

Bảng 4.6. Kết quả với dữ liệu kiểm thử DUC-2003

	RG-1	RG-2	RG-L
Bộ lọc 1	15.39	3.72	14.31
Bộ lọc 2	14.38	3.67	13.36
Bộ lọc 3	**16.69**	**4.64**	**15.27**
Bộ lọc 4	14.83	3.87	13.84
Bộ lọc 5	16.15	4.12	14.99

Bảng 4.7. Kết quả với dữ liệu kiểm thử DUC-2004

	RG-1	RG-2	RG-L
Bộ lọc 1	12.89	3.22	11.78
Bộ lọc 2	12.39	3.06	11.30
Bộ lọc 3	**14.23**	**3.73**	**12.93**
Bộ lọc 4	12.63	3.26	11.63
Bộ lọc 5	13.63	3.31	12.39

Kết quả trên các tập kiểm thử cho thấy, độ chính xác tốt nhất đạt được khi sử dụng bộ lọc 3. Tức là tập đặc trưng 1-grams, 2-grams và 3-grams cho kết quả tốt nhất. Kết quả trên bộ dữ liệu kiểm thử Gigaword đạt cao nhất, sau đó đến bộ dữ liệu DUC-2003 và DUC-2004. Nguyên nhân là do sử dụng dữ liệu Gigaword để huấn luyện mô hình.

Tôi so sánh kết quả trên với kết quả khi chạy mô hình words-lvt2k-1sent (GRU với cơ chế Attention) của nhóm tác giả Nallapati[19]:

Bảng 4.8. Kết quả mô hình words-lvt2k-1sent

	RG-1	RG-2	RG-L
Gigaword	16.59	4.26	15.74
DUC-2003	6.41	1.11	6.12
DUC-2004	5.69	0.81	5.47

Kết quả so sánh trên cho thấy hiệu quả rõ ràng của mô hình CNN khi áp dụng vào mạng GRU với cơ chế Attention.

Bảng 4.9. Ví dụ đầu ra với bộ dữ liệu Gigaword

Ví dụ 1	the sri lankan government on wednesday announced the closure of government schools with immediate effect as a military campaign against tamil separatists escalated in the north of the country .
Câu tóm tắt	sri lanka closes schools as war escalates
Câu sinh ra	sri lanka announces UNK of schools
Ví dụ 2	police arrested five anti-nuclear protesters thursday after they sought to disrupt loading of a french antarctic research and supply vessel , a spokesman for the protesters said .
Câu tóm tắt	protesters target french research ship
Câu sinh ra	french police arrest five protesters

Ví dụ 3	factory orders for manufactured goods rose #.# percent in september , the commerce department said here thursday .
Câu tóm tắt	us september factory orders up #.# percent
Câu sinh ra	us factory orders up #.# percent in september
Ví dụ 4	croatian president franjo tudjman said friday croatian and serb negotiators would meet saturday to thrash out an agreement on the last serb-held area in croatia , under a deal reached at us-brokered talks .
Câu tóm tắt	rebel serb talks to resume saturday : tudjman by peter UNK
Câu sinh ra	croatia and croatia to resume talks
Ví dụ 5	israel prepared sunday for prime minister yitzhak rabin 's state funeral which will be attended by a host of world leaders , including us president bill clinton and the jordanian and egyptian heads of state .
Câu tóm tắt	israel prepares jerusalem state funeral for rabin
Câu sinh ra	israel prepares for UNK state funeral

Kết quả cho thấy câu sinh ra gần giống với câu tóm tắt, tuy nhiên một số câu sinh ra gặp phải các vấn đề như:

- Ngữ pháp không đúng
- Vấn đề các từ hiếm (từ UNK) xuất hiện trong câu
- Vấn đề lặp từ

Đây cũng là những vấn đề thách thức đặt ra cho các nhà nghiên cứu tìm phương pháp giải quyết.

4.3.2. Bộ dữ liệu CNN/Daily Mail

Tôi chạy thí nghiệm với kiểu bộ lọc 3: sử dụng ba bộ lọc với kích thước tương ứng 200 – 250 – 300.

Do cấu hình máy huấn luyện hạn chế nên tôi dừng lại quá trình huấn luyện ở epoch 10 để kiểm tra kết quả.

Bảng 4.10. Kết quả với bộ dữ liệu CNN/Daily Mail

	RG-1	RG-2	RG-L
Kết quả	18.39	2.95	13.76

Bảng 4.11. Ví dụ đầu ra với bộ dữ liệu CNN/Daily Mail

Ví dụ 1	the Michigan has decided to proceed with a screening of the film " American Sniper " despite objections from some students more than 200 students signed a petition asking the school not to show the movie as part of UMix , a series of social events the university stages for students Bradley Cooper was nominated for an Oscar for his portrayal of Kyle , a Navy seal and the most lethal sniper in American military history Kyle was fatally shot at a Texas shooting range in 2013 some students believed the movie 's depiction of the Iraq War reflected negatively on the Middle East and people from that region Michigan 's Detroit metropolitan area is home to the nation 's largest Arab - American population but there was a backlash to the decision to yank the movie , and a counter-petition asked school officials to reconsider on wednesday , E. Royster Harper , Michigan 's vice president for student life , said in a statement that " it was a mistake to cancel the showing of the movie ' American Sniper ' on campus as part of a social event for students " and that the show will go on " the initial decision to cancel the movie was not consistent with the high value the Michigan places on freedom of expression and our respect for the right of students to make their own choices in such matters , " the statement said UMix will offer a screening of the family - friendly " Paddington " for those who would rather not attend " American Sniper " the announcement drew praise from Michigan head football coach Jim Harbaugh .
Văn bản tóm tắt	some *complained* about the film 's depiction of the Iraq War a petition asked the university not to show the Bradley Cooper film
Văn bản sinh ra	the video was posted on the website of the UNK Academy in Michigan the video shows the school students at the school in Michigan
Ví dụ 2	Tokyo a bizarre and alarming discovery is raising concerns in Japanese about the potential for terrorism involving drones a drone carrying traces

	of a radioactive material was found on the rooftop of Japanese 's equivalent to the White House on wednesday , police and government officials said the discovery came on the same day a Japanese court approved a government plan to restart two reactors at the Sendai nuclear power plant in Kagoshima prefecture , more than four years after the Fukushima Daiichi nuclear disaster prime minister Abe 's push to restart the reactors is unpopular among many Japanese , who view nuclear energy as too dangerous a staff member spotted the drone wednesday morning on the roof of Abe 's residence , Tokyo Metropolitan Police said dozens of police investigators were dispatched to the roof to investigate the origin of the drone , which had four propeller and was 50 centimeters (20 inches) wide police say the drone was equipped with a small camera , smoke flares and a plastic bottle containing small traces of a radioactive material believed to be cesium , a common byproduct of nuclear reactors cesium was also discovered in areas around the failed Fukushima Daiichi nuclear plant after its 2011 meltdown investigators suspect the cesium was placed in the bottle the amount inside is not immediately harmful to humans chief cabinet secretary Suga said the discovery is raising concerns about terrorism " there might be terrorism attempts in the future at the Olympics and G7 Summit using drones , " Suga said " so we need to examine and review continuously the way small unmanned vehicles like drones should be operated and how to cope with the threat of terrorism from drones the government will do all that we can to prevent terrorism " Japanese law restricts drone flights around airports to prevent problems with aircraft , but there are no flight restrictions for most of Tokyo , including the prime minister 's residence and local and federal government buildings Abe was not in his office at the time he is in Indonesia , attending the Asian-African Conference CNN 's Elizabeth Joseph , Joshua Berlinger and Josh Levs contributed to this report .
Văn bản tóm tắt	the drone is *sparking* terrorism concerns , authorities say it was equipped with a bottle containing radioactive material it was discovered as a court approved a plan to restart two Japanese nuclear reactors
Văn bản sinh ra	the device was used by the White House in UNK , the White House it is believed to have been caused by a drone strikes in the world
Ví dụ 3	think it 's hard to redeem your miles for an airline award ticket ? it depends on which airline rewards program you 've chosen , which route

	you 're flying and when you book your ticket , according to a new Consumer Reports study of 70 million passenger trips over the past two years the magazine collected statistics comparing award - seat availability for the five biggest American airlines on domestic routes the top performer was Southwest Airlines , which offered the most award tickets , 11.9 million , and the highest percentage of award tickets -- 11.5 % of 103.1 million total passenger seats " the high number of award tickets is directly related to Southwest Airlines 's unique combination of ' every seat is an Award Seat , ' no blackout dates , points that do n't expire , and a route map that reaches more than 90 different destinations in the American and beyond , making us the largest domestic carrier in the American , " Southwest Airlines spokesperson Thais Conway Hanson told CNN " unlike other carriers , we also do n't charge fees for close - in bookings or penalize you for canceling your trip if something else comes up " at the bottom of the list was JetBlue , which offered the lowest percentage of award seats and the fewest number of award tickets of the five biggest American airlines : 892,000 one - way passenger tickets , or 4.5 % of its total 19.7 million American seats (JetBlue only operates in 10 of the top 25 markets included in the study) many JetBlue customers fly the airline only once or twice per year , making it hard to accumulate miles , an airline spokesman told the magazine by not allowing miles to expire anymore , the airline says customers will be able to eventually redeem them Delta came in second place with 5.6 million American award seats ; United ranked third with 5 million American award tickets ; and American Airlines was fourth with 3.5 million American award seats what are the world 's safest airlines ? remember that award tickets are n't actually free the cost of miles is built into everything you buy that 's earning you miles , and the airlines profit from you not using your miles at all so it behooves consumers to book award travel carefully on average , nearly 10 % of passengers on the five airlines analyzed by Consumer Reports flew on domestic award tickets , but some of them were n't getting the best value for their miles while many American fliers redeemed miles on American Airlines flights from Los Angeles to San Francisco , the cheapest average fare on that route was just over $ 100 -- not worth the 12,500 to 30,000 miles needed for an award ticket , Consumer Reports says better to use them on American Airlines ' route between New York and San Francisco or Delta 's route between Chicago and Los Angeles , which are generally more expensive than that Los Angeles - San Francisco route , according to Consumer Reports '

	calculations while award - seat availability is important , it may not matter as much as passengers ' overall satisfaction with an airline Southwest Airlines had the highest customer satisfaction score (86) , followed by JetBlue (85) , Delta (70) , American (66) and United (63) and do n't forget the fees Southwest Airlines does n't charge any fees , while other airlines tack on fees for checking bags , booking by phone , changing plans and more .
Văn bản tóm tắt	Southwest Airlines tops Consumer Reports ' survey , with the most seats available JetBlue is at the bottom of the list but ranks high in customer satisfaction
Văn bản sinh ra	UNK UNK , UNK , UNK , UNK , UNK , UNK , UNK and UNK are the most popular airline 's top - ranked airline 's top - ranked airline UNK UNK UNK UNK UNK UNK , UNK and UNK

Kết quả cho thấy, đoạn tóm tắt đưa ra chưa thể hiện đúng nội dung tóm tắt như người dùng, chúng còn sai về ngữ pháp và gặp nhiều vấn đề khác tương tự như với bộ dữ liệu Gigaword.

Kết luận

Luận văn là một nghiên cứu cho bài toán tóm tắt văn bản theo hướng tóm lược ý, thực nghiệm tiến hành trên dữ liệu tiếng Anh.

Nghiên cứu đã có kết quả bước đầu cho bài toán tóm tắt văn bản. Luận văn đã trình bày một số vấn đề chính sau:

- Tìm hiểu tổng quan về tóm tắt văn bản và đi sâu vào tóm tắt tóm lược.
- Trình bày hiểu biết về các mô hình mạng trong học sâu như mạng nơ-ron đa lớp, mạng LSTM, mạng GRU, mạng nơ-ron tích chập.
- Đề xuất mô hình dựa trên mạng nơ-ron tích chập và mạng GRU kèm theo cơ chế attention.
- Tiến hành thử nghiệm với hai bộ dữ liệu khác biệt với các cấu hình mạng CNN khác nhau. Kết quả cho thấy hiệu quả rõ ràng của mô hình đề xuất so với mô hình words-lvt2k-1sent của nhóm tác giả R Nallapati [19].

Mặc dù đã cố gắng và nỗ lực, nhưng do thời gian nghiên cứu và trình độ bản thân có hạn cùng với cấu hình máy chạy chưa đủ mạnh nên luận văn chưa đạt được kết quả như mong muốn.

Trong tương lai, tôi tiếp tục hướng nghiên cứu dùng các mô hình Deep learning mới cho bài toán tóm tắt văn bản theo hướng tóm lược:

- Sử dụng Cơ chế bao phủ [19], [23]: Sự lặp lại từ có thể được tính toán bằng sự tăng lên và liên tục chú ý tới một từ cụ thể.
- Sử dụng mạng Pointer [23]: Các bản tóm tắt cần sao chép hoặc chứa một lượng các từ xuất hiện trong văn bản nguồn.
- Sử dụng các phương pháp học tăng cường [24]: dựa vào độ đo ROUGE để định nghĩa hàm lỗi.

Tài liệu tham khảo

1. Ani Nenkova and Kathleen McKeown, Automatic Summarization, Foundations and Trends in Information Retrieval, Vol. 5: No. 2–3, pp 103-233.
2. Alexander M. Rush and Sumit Chopra and Jason Weston (2015), A Neural Attention Model for Abstractive Sentence Summarization, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 379-389.
3. Sumit Chopra and Michael Auli and Alexander M. Rush (2016), Abstractive Sentence Summarization with Attentive Recurrent Neural Networks, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, pp. 93-98.
4. Qingyu Zhou and Nan Yang and Furu Wei and Ming Zhou (2017), Selective Encoding for Abstractive Sentence Summarization, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1095-1104.
5. Yoon Kim (2014), Convolutional Neural Networks for Sentence Classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, pp. 1746-1751
6. Nal Kalchbrenner and Edward Grefenstette and Phil Blunsom (2014), A Convolutional Neural Network for Modelling Sentences, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, pp. 655-665.
7. Yoon Kim and Yacine Jernite and David Sontag, Alexander M. Rush (2016), Character-Aware Neural Language Models, Proceedings of the Thirtieth Conference on Artificial Intelligence, Phoenix, Arizona, USA.
8. Jason Lee and Kyunghyun Cho and Thomas Hofmann (2017), Fully Character-Level Neural Machine Translation without Explicit, Transactions of the Association for Computational Linguistics, pp. 365-378.
9. Dzmitry Bahdanau and Kyunghyun Cho and Yoshua Bengio (2015), Neural Machine Translation by Jointly Learning to Align and Translate, International Conference on Learning Representations.

10. Ilya Sutskever and Oriol Vinyals and Quoc V. Le (2014), Sequence to Sequence Learning with Neural Networks, Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, Quebec, Canada, pp. 3104-3112.
11. Thang Luong and Hieu Pham and Christopher D. Manning (2015), Effective Approaches to Attention-based Neural Machine Translation, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 1412-1421.
12. Kyunghyun Cho and Bart van Merrienboer and Caglar Gulcehre and Dzmitry Bahdanau and Fethi Bougares and Holger Schwenk and Yoshua Bengio (2014), Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, pp.1724-1734.
13. Junyoung Chung and KyungHyun Cho and Yoshua Bengio (2014), Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, Advances in Neural Information Processing Systems 2014 Deep Learning and Representation Learning Workshop.
14. Rupesh Kumar Srivastava and Klaus Greff and Jurgen Schmidhuber (2015), Training Very Deep Networks, Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, Quebec, Canada.
15. Kyunghyun Cho and Bart van Merrienboer and Dzmitry Bahdanau, Yoshua Bengio (2014), On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, pp. 103-111.
16. Lin, Chin-Yew (2004), ROUGE: a Package for Automatic Evaluation of Summaries, Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, pp. 74-81.
17. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin (2017), Convolutional Sequence to Sequence Learning, Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia.

18. Ian Goodfellow and Yoshua Bengio, and Aaron Courville (2016), Deep Learning, MIT Press.
19. R Nallapati, B Zhou, C Gulcehre, B Xiang (2016), Abstractive Text Summarization using Sequence-to-Sequence RNNs and Beyond, The SIGNLL Conference on Computational Natural Language Learning, pp. 280-290.
20. Jianpeng Cheng and Mirella Lapata (2016), Neural summary by extracting sentences and words, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, pp. 484-494.
21. Alex Graves (2012), Supervised Sequence Labelling with Recurrent Neural Networks, Studies in Computational Intelligence, Springer.
22. N Moratanch, S Chitrakala (2016), A survey on abstractive text summarization, International Conference on Circuit, Power and Computing Technologies.
23. Abigail See, Peter J. Liu, Christopher D. Manning (2017), Get To The Point: Summarization with Pointer-Generator Networks, Annual Meeting of the Association for Computational Linguistics, pp. 1073-1083.
24. Romain Paulus, Caiming Xiong, Richard Socher (2018), A Deep Reinforced Model for Abstractive Summarization, 6th International Conference on Learning Representations.
25. Nguyễn Viết Hạnh (2018), Nghiên cứu tóm tắt văn bản tự động và ứng dụng, Luận văn thạc sĩ, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.