Điện toán thần kinh 452 (2021) 48-62

Danh sách nội dung có sẵn tại ScienceDirect

Điện toán thần kinh

trang chủ tạp chí: www.elsevier.com/locate/neucom

Đánh giá về cơ chế chú ý của học sâu

Triệu Dương Ngưu [một], Quốc Cường Trung [một], Hui Yu [b],

aKhoa Khoa học và Công nghệ Máy tính, Đại học Hải dương Trung Quốc, Thanh Đảo 266100, Trung Quốc
bTrường Công nghệ Sáng tạo, Đại học Portsmouth, Portsmouth PO1 2DJ, Vương quốc Anh

thông tin bài viết

Lịch sử bài viết:
Nhận vào ngày 7 tháng 2 năm 2021
Sửa đổi ngày 23 tháng 3 năm 2021
Được chấp nhận ngày 29 tháng 3 năm 2021
Có sẵn trực tuyến vào ngày 1 tháng 4 năm 2021
Được truyền đạt bởi Zidong Wang

Từ khóa:
Cơ chế chú ý
Học kĩ càng
Mạng thần kinh tái phát (RNN)
Mạng thần kinh chuyển đổi (CNN)
Bộ mã hóa-giải mã
Mô hình chú ý thống nhất
Ứng dụng thị giác máy tính
Ứng dụng xử lý ngôn ngữ tự nhiên

trừu tượng

Sự chú ý được cho là đã trở thành một trong những khái niệm quan trọng nhất trong lĩnh vực học sâu. Nó được lấy cảm hứng từ hệ thống sinh học của con người có xu hướng tập trung vào các phần đặc biệt khi xử lý lượng lớn thông tin. Với sự phát triển của mạng lưới thần kinh sâu, cơ chế chú ý đã được sử dụng rộng rãi trong các lĩnh vực ứng dụng đa dạng. Bài viết này nhằm mục đích cung cấp một cái nhìn tổng quan về các mô hình chú ý tiên tiến được đề xuất trong những năm gần đây. Để hiểu rõ hơn về cơ chế chú ý, chúng tôi xác định một mô hình thống nhất phù hợp với hầu hết các cấu trúc chú ý. Mỗi bước của cơ chế chú ý được triển khai trong mô hình đều được mô tả chi tiết. Hơn nữa, chúng tôi phân loại các mô hình chú ý hiện có theo bốn tiêu chí: mức độ chú ý, dạng tính năng đầu vào, biểu diễn đầu vào và biểu diễn đầu ra. Ngoài ra, chúng tôi tóm tắt các kiến trúc mạng được sử dụng cùng với cơ chế chú ý và mô tả một số ứng dụng điển hình của cơ chế chú ý. Cuối cùng, chúng tôi thảo luận về khả năng diễn giải mà sự chú ý mang lại cho việc học sâu và trình bày các xu hướng tiềm năng trong tương lai của nó.

1. Giới thiệu

Chú ý là một chức năng nhận thức phức tạp không thể thiếu đối với con người [1,2]. Một đặc tính quan trọng của nhận thức là con người không có xu hướng xử lý toàn bộ thông tin cùng một lúc. Thay vào đó, con người có xu hướng tập trung có chọn lọc vào một phần thông tin khi nào và ở đâu cần thiết, nhưng đồng thời lại bỏ qua những thông tin có thể nhận biết được khác. Ví dụ, con người thường không nhìn thấy tất cả các cảnh từ đầu đến cuối khi nhận thức bằng mắt, mà thay vào đó, quan sát và chú ý đến những phần cụ thể khi cần thiết. Khi con người nhận thấy một cảnh thường có thứ họ muốn quan sát ở một phần nào đó, họ sẽ học cách tập trung vào phần đó khi những cảnh tương tự xuất hiện trở lại và tập trung chú ý hơn vào phần hữu ích. Đây là phương tiện để con người nhanh chóng lựa chọn thông tin có giá trị cao từ lượng thông tin khổng lồ sử dụng tài nguyên xử lý hạn chế. Cơ chế chú ý cải thiện đáng kể hiệu quả và độ chính xác của việc xử lý thông tin nhận thức.

Cơ chế chú ý của con người có thể được chia thành hai phân loại theo cách thức tạo ra nó [3]. Loại đầu tiên là sự chú ý vô thức từ dưới lên, được gọi là sự chú ý dựa trên độ mặn, được thúc đẩy bởi các kích thích bên ngoài. Ví dụ, mọi người có nhiều khả năng nghe thấy giọng nói lớn hơn trong một cuộc trò chuyện. Nó tương tự như cơ chế gộp tối đa và gating [4,5] trong học sâu, chuyển các giá trị phù hợp hơn (tức là các giá trị lớn hơn) sang bước tiếp theo. Loại thứ hai là sự chú ý có ý thức từ trên xuống, được gọi là sự chú ý tập trung. Sự chú ý tập trung đề cập đến sự chú ý có mục đích được xác định trước và dựa vào các nhiệm vụ cụ thể.

Nó cho phép con người tập trung sự chú ý vào một đối tượng nhất định một cách có ý thức và tích cực. Hầu hết các cơ chế chú ý trong deep learning đều được thiết kế theo các nhiệm vụ cụ thể sao cho hầu hết đều là sự chú ý tập trung. Cơ chế chú ý được giới thiệu trong bài viết này thường đề cập đến sự chú ý tập trung ngoại trừ những trường hợp đặc biệt.
các câu lệnh.

Như đã đề cập ở trên, cơ chế chú ý có thể được sử dụng như một sơ đồ phân bổ nguồn lực, đây là phương tiện chính để giải quyết vấn đề quá tải thông tin. Trong trường hợp sức mạnh tính toán hạn chế, nó có thể xử lý thông tin quan trọng hơn với nguồn lực tính toán hạn chế. Do đó, một số nhà nghiên cứu chú ý đến lĩnh vực thị giác máy tính. [6] đã đề xuất một mô hình chú ý trực quan dựa trên độ nổi bật để trích xuất các đặc điểm hình ảnh cấp thấp cục bộ để có được một số vùng nổi bật tiềm năng. Trong khu vực mạng nơ-ron, [7] đã sử dụng cơ chế chú ý trên mô hình mạng nơ-ron hồi quy để phân loại hình ảnh. Bahdanau và cộng sự. [8] đã sử dụng cơ chế chú ý để thực hiện đồng thời việc dịch và căn chỉnh các tác vụ dịch máy. Sau đó, cơ chế chú ý đã trở thành một thành phần ngày càng phổ biến của kiến trúc thần kinh và được áp dụng cho nhiều tác vụ khác nhau, chẳng hạn như chú thích hình ảnh.

Tác giả tương ứng.
Địa chỉ email: nnniuzy@163.com (Z. Niu), gqzhong@ouc.edu.cn (G. Zhong), hui.yu@port.ac.uk (H.Yu).

https://doi.org/10.1016/j.neucom.2021.03.091
0925-2312/

thế hệ [9,10], phân loại văn bản [11,12], dịch máy [13–16], nhận dạng hành động [17–20], phân tích dựa trên hình ảnh [21,22], nhận dạng giọng nói [23–25], khuyến nghị [26,27], và đồ thị [28,29]. Hình 1 cho thấy tổng quan về một số điển hình phư ơ ng pháp chú ý, đư ợc mô tả chi tiết trong Phần 3.

Ngoài việc cung cấp các cải tiến về hiệu suất, cơ chế chú ý cũng có thể đư ợc sử dụng như một công cụ để giải thích hành vi kiến trúc thần kinh khó hiểu. Trong những năm gần đây, mạng nơ -ron đã đạt đư ợc thành công lớn trong lĩnh vực tài chính [31], vật chất [32], khí tư ợng học [33–36], y tế [37–41], lái xe tự động [42], tư ơ ng tác giữa ngư ời và máy tính [43–45], phân tích hành vi và hành động [46,47], các ngành [48] và phát hiện cảm xúc [49,50], như ng thiếu khả năng diễn giải đang phải đối mặt với cả vấn đề thực tế và đạo đức [51]. Khả năng diễn giải của deep learning là một vấn đề xa. Mặc dù liệu cơ chế chú ý có thể đư ợc sử dụng như một phư ơ ng pháp đáng tin cậy để giải thích mạng lư ới sâu vẫn còn gây tranh cãi vấn đề [52,53], nó có thể cung cấp lời giải thích trực quan cho một vấn đề nhất định mức độ [54–57]. Ví dụ: Hình 2 cho thấy một ví dụ về trực quan hóa trọng lư ợng sự chú ý.

Cuộc khảo sát này đư ợc cấu trúc như sau. Trong Phần 2, chúng tôi giới thiệu một mô hình nổi tiếng đư ợc đề xuất bởi [8] và xác định sự chú ý chung ngư ời mẫu. Phần 3 mô tả việc phân loại các mô hình chú ý. Phần 4 tóm tắt các kiến trúc mạng kết hợp với cơ chế chú ý Phần 5 trình bày chi tiết về cách sử dụng sự chú ý trong các nhiệm vụ thị giác máy tính (CV) và xử lý ngôn ngữ tự nhiên (NLP) khác nhau. Trong Phần 6, chúng tôi thảo luận về khả năng diễn giải sự chú ý mang lại cho mạng lư ới thần kinh. Trong Phần 7, chúng tôi mổ xẻ mở những thách thức, xu hư ớng hiện tại và những cách thức đổi mới trong sự chú ý cơ chế. Trong Phần 8, chúng tôi kết luận bài viết này.

2. Cơ chế chú ý

Trong phần này, trư ớc tiên chúng tôi mô tả một kiến trúc dịch máy nổi tiếng bằng cách sử dụng sự chú ý đư ợc giới thiệu bởi [8] và coi nó như một ví dụ để xác định một mô hình chú ý chung.

2.1. Một ví dụ về mô hình chú ý: RNNsearch

Bahdanau và cộng sự. [8] đề xuất một mô hình chú ý, RNNsearch, lần đầu tiên áp dụng cơ chế chú ý vào tác vụ dịch máy. RNNsearch bao gồm mạng nơ -ron tái phát hai chiều (BiRNN) [58] làm bộ mã hóa và bộ giải mã mô phỏng việc tìm kiếm thông qua câu nguồn khi giải mã một bản dịch, như đư ợc minh họa trong Hình 3.

Bộ mã hóa tính toán chú thích ðh1; ... hT Þ đó là trạng thái ẩn của BiRNN dựa trên chuỗi đầu vào ðx1; ... xT Þ:

$$ðh1; ... ; hT Þ ¼ BiRNNðx1; ... ; xTÞ : \quad ð1Þ$$

Một thiếu sót của RNN thông thư ờng là chúng chỉ sử dụng bối cảnh trư ớc đó. BiRNN có thể đư ợc huấn luyện bằng cách sử dụng tất cả các dữ liệu đầu vào có sẵn thông tin trong quá khứ và tư ơ ng lai của một khung thời gian cụ thể. cụ thể- về cơ bản, như trong Hình 3, các trạng thái ẩn fh1 ; ··· ; hTg và fh1 ; ht· g· ; đư ợc trích xuất bằng RNN tiến và RNN lùi tư ơ ng ứng. Sau đó, bộ mã hóa thu đư ợc chú thích cho một từ xi bằng cách nối trạng thái ẩn về phía trư ớc xin chào và cái ngư ợc lại xin chào , > tức là xin chào ¼ ½hi Þ; CHÀO .

Bộ giải mã bao gồm một khối chú ý và mạng thần kinh tái phát (RNN). Chức năng của khối chú ý là tính toán vectơ ngữ cảnh c đại diện cho mối quan hệ ngữ cảnh giữa ký hiệu đầu ra hiện tại và mỗi số hạng của toàn bộ trình tự đầu vào. Tại mỗi bư ớc thời gian t, vectơ ngữ cảnh ct đư ợc tính dư ới dạng tổng có trọng số của các chú thích hj này:

$$ct ¼ \sum_{j¼1}^{T} tajhj : \quad ð2Þ$$

Trọng số chú ý atj của mỗi chú thích hj đư ợc tính bằng

$$etj ¼ aðst1; hjÞ; \quad ð3Þ$$

Và

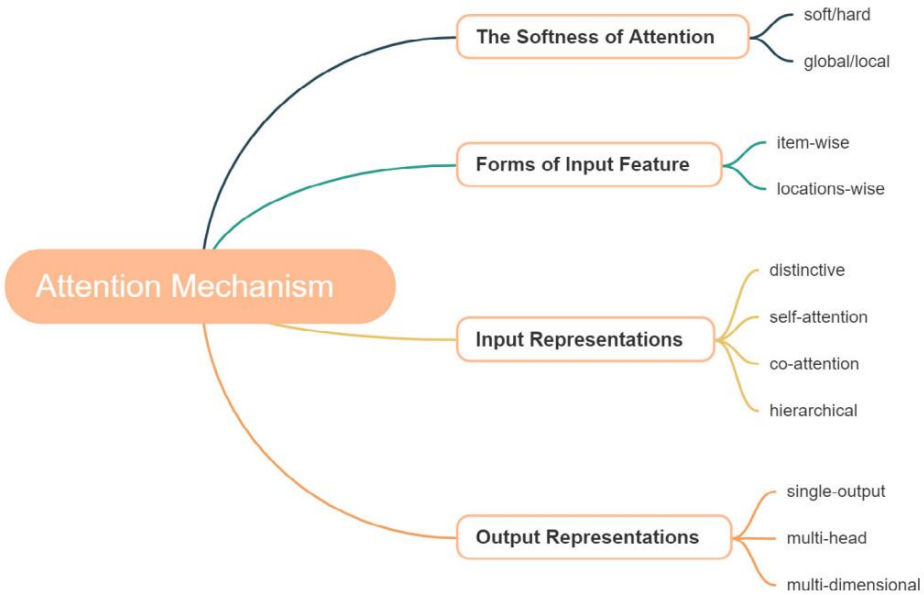
$$tại ¼ \frac{kinh nghiệm}{\sum_{k¼1}^{T} kinh nghiệm} ; \quad ð4Þ$$

trong đó a là một hàm có thể học đư ợc và nó phản ánh tầm quan trọng của chú thích hj sang trạng thái ẩn tiếp theo st theo trạng thái st1.

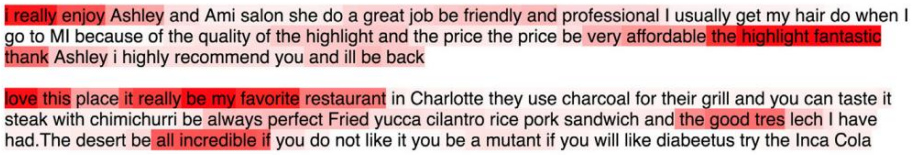
Sau đó, RNN xuất ra ký hiệu yt có khả năng xảy ra cao nhất tại bư ớc hiện tại:

$$pðytjy1; ... ; yt1; xÞ ¼ RNNðctÞ: \quad ð5Þ$$

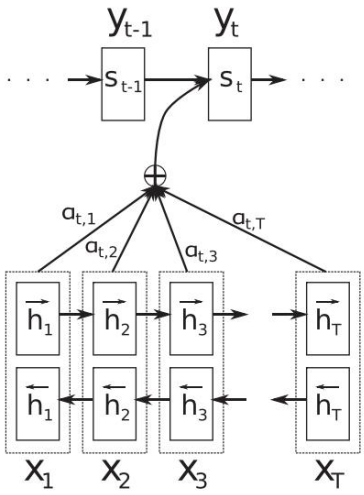
Bằng cách này, thông tin của câu nguồn có thể đư ợc phân bổ thành toàn bộ chuỗi thay vì mã hóa tất cả thông tin thành một vectơ có độ dài cố định thông qua bộ mã hóa, trong khi



Hình 1. Một số cách tiếp cận điển hình đối với cơ chế chú ý.



Hình 2. Sơ đồ nhiệt của các bài đánh giá Yelp 5 sao từ [30]. Màu sắc đậm hơn cho thấy mức độ chú ý cao hơn.



Hình 3. Minh họa một bước giải mã trong máy thần kinh dựa trên sự chú ý dịch [8].

bộ giải mã có thể lấy nó một cách có chọn lọc ở mỗi bước thời gian. Công thức này cho phép mạng lưới thần kinh tập trung vào các yếu tố liên quan của đầu vào khác với những phần không liên quan.

2.2. Mô hình chú ý thống nhất

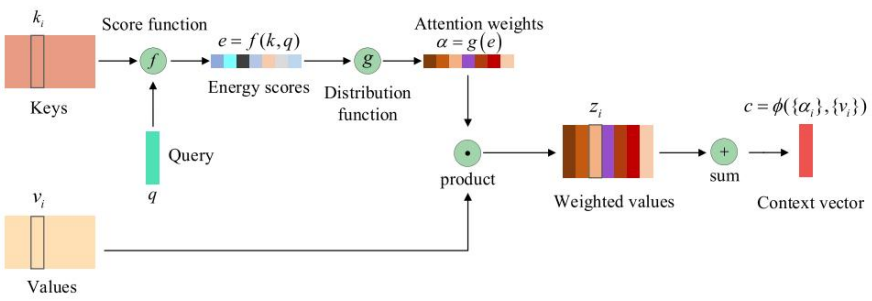
Sau khi [8] áp dụng cơ chế chú ý vào các tác vụ dịch máy, các biến thể mô hình chú ý được sử dụng trong các ứng dụng khác nhau tên miền đã phát triển nhanh chóng. Nói chung, quá trình thực hiện cơ chế chú ý có thể được chia thành hai bước: một là tính toán sự phân bố sự chú ý trên thông tin đầu vào, và cách khác là tính toán vectơ ngữ cảnh theo phân phối sự chú ý. Hình 4 thể hiện mô hình chú ý thống nhất chúng tôi đã xác định, bao gồm phần cốt lõi được chia sẻ bởi hầu hết mô hình chú ý được tìm thấy trong các tài liệu được khảo sát.

Khi tính toán phân bổ sự chú ý, mạng lưới thần kinh đầu tiên mã hóa tính năng dữ liệu nguồn dưới dạng K, được gọi là khóa. K có thể được thể hiện dưới nhiều hình thức khác nhau tùy theo nhiệm vụ cụ thể và cấu trúc thần kinh. Ví dụ, K có thể là đặc điểm của một chứng chỉ vùng hình ảnh, phần nhúng từ của tài liệu hoặc trạng thái ẩn của RNN, như xảy ra với các chú thích trong RNNsearch. Ngoài ra, thông thường cần phải giới thiệu một vectơ biểu diễn liên quan đến nhiệm vụ q, truy vấn, giống như trạng thái ẩn trước đó của đầu ra st1 trong RNNsearch. Trong một số trường hợp, q cũng có thể ở dạng ma trận [16] hoặc hai vectơ [59] tùy theo nhiệm vụ cụ thể.

Sau đó, mạng lưới thần kinh tính toán mối tương quan giữa các truy vấn và khóa thông qua hàm tính điểm f (còn gọi là hàm năng lượng hàm [61] hoặc hàm tương thích [62]) để thu được năng lượng điểm e phản ánh tầm quan trọng của truy vấn liên quan đến khóa trong việc quyết định đầu ra tiếp theo:

e ¼ fðq;KÞ: ð6Þ

Hàm tính điểm f là một phần quan trọng của mô hình chú ý bởi vì nó xác định cách các khóa và truy vấn được khớp hoặc kết hợp. Trong Bảng 1, chúng tôi liệt kê một số hàm tính điểm phổ biến. Hai nhất cơ chế chú ý thường được sử dụng là sự chú ý bổ sung (như mô hình căn chỉnh trong RNNsearch) [8] và tính toán chú ý nhân (sản phẩm chấm) ít tốn kém hơn [14]. britz et al. [15] đã thực hiện so sánh thực nghiệm giữa hai điểm số này chức năng. Trong thí nghiệm của họ trên WMT'15 Tiếng Anh!Tiếng Đức nhiệm vụ, họ phát hiện ra rằng các cơ chế chú ý cộng gộp được tham số hóa hơi như ng luôn hoạt động tốt hơn cơ chế nhân một. Hơn nữa, Vaswani et al. [16] đã đề xuất một biến thể của sự chú ý nhân lên bằng cách thêm hệ số tỷ lệ 1ffiffiffiffi $\frac{}{dk}$ p , dk ở đâu kích thước của các phím. Trong khi đối với các giá trị nhỏ của dk, hai cơ chế hoạt động tương tự nhau, thì sự chú ý cộng gộp sẽ vượt trội hơn sự chú ý nhân nhiều mà không cần điều chỉnh tỷ lệ đối với các giá trị lớn hơn của dk. Cũng, Lương và cộng sự. [14] trình bày sự chú ý chung, concat chú ý, và sự chú ý dựa trên vị trí. Sự chú ý chung mở rộng khái niệm về sự chú ý nhân lên bằng cách giới thiệu một ma trận có thể học được tham số W, có thể được áp dụng cho các khóa và truy vấn có cách biểu diễn khác nhau. Sự chú ý của Concat nhằm mục đích lấy được khớp biểu diễn các khóa và truy vấn thay vì so sánh chúng. Nó tương tự như sự chú ý bổ sung ngoại trừ việc tính toán q và K riêng biệt. Trong sự chú ý dựa trên vị trí, điểm căn chỉnh chỉ được tính toán từ trạng thái ẩn mục tiêu. Nói cách khác, năng lượng điểm chỉ liên quan đến q khác với K. Ngược lại, sự tự chú ý



Hình 4. Kiến trúc của mô hình chú ý thống nhất.

Bảng 1

Tóm tắt hàm tính điểm f. Ở đây, k là một phần tử của K; v; b;W; W1; W2 có thể học được tham số, dk là thứ nguyên của vectơ đầu vào. Hành động này là một sự kích hoạt phi tuyến tính chức năng, chẳng hạn như tanh và ReLU.

Tên	phương trình	Tham chiếu
phụ gia	fðq; kÞ ¼ vTactðW1k þ W2q þ bÞ	[15]
Nhân (sản phẩm chấm)	fðq; kÞ ¼ qTk	[15]
nhân theo tỷ lệ	$f(q,k) = \frac{qTk}{dk}$	[16]
Tổng quan	p fðq; kÞ ¼ qTWk	[14]
Concat	fðq; kÞ ¼ vTact Wð ½q; k þ bÞ	[14]
Dựa trên địa điểm	fðq; kÞ ¼ fðqÞ	[14]
Sự tương đồng	$f(q,k) = \frac{qk}{\|k\|\|q\|}$	[60]

[63] chỉ được tính toán dựa trên K mà không cần q. Hơn nữa, Graves et al. [60] đã trình bày một mô hình so sánh độ tương tự giữa K và q, dựa trên độ tương tự cosine.

Sau đó, điểm năng lượng e được ánh xạ tới trọng số chú ý a thông qua hàm phân phối chú ý g:

một ¼ gðeÞ: ð7Þ

Hàm phân phối g tương ứng với softmax trong RNNsearch, chuẩn hóa tất cả các điểm năng lượng thành một xác suất phân bổ. Ngoài softmax, một số nhà nghiên cứu đã thử chức năng phân phối khác. Hạn chế của hàm softmax là phân bố xác suất thu được luôn có hỗ trợ đầy đủ, tức là softmaxðzÞ > 0 cho mọi số hạng của z. Đây là một bất lợi trong các ứng dụng mong muốn có sự phân bố xác suất thưa thớt, trong trường hợp nào [64] đề xuất mức tối đa có thể chỉ định chính xác bằng 0 xác suất đối với một số biến đầu ra của nó. Ngoài ra, [65] đã sử dụng một hàm phân phối khác, sigmoid logistic, được chia tỷ lệ điểm năng lượng nằm trong khoảng từ 0 đến 1. Họ cũng so sánh sigmoid với softmax trong các thí nghiệm của họ và kết quả cho thấy rằng sigmoid chức năng thực hiện tốt hơn hoặc kém hơn trong các nhiệm vụ khác nhau.

Khi mạng nơ-ron tính toán các vectơ ngữ cảnh, nó thường cần thiết để giới thiệu một biểu diễn tính năng dữ liệu mới V, được gọi là giá trị. Mỗi phần tử của V tương ứng với một và chỉ một phần tử của K. Trong nhiều kiến trúc, cả hai đều là sự biểu diễn giống nhau của dữ liệu đầu vào, giống như các chú thích trong RNNsearch. Dựa trên công trình trước đây [66–68], Daniluk et al. [69] đưa ra giả thuyết rằng việc sử dụng quá nhiều các biểu diễn này gây khó khăn cho việc huấn luyện mô hình, vì vậy họ đã đề xuất sửa đổi cơ chế chú ý để phân tách các chức năng này một cách rõ ràng. Họ sử dụng các cách biểu diễn khác nhau của đầu vào để tính toán mức độ chú ý phân phối và thông tin theo ngữ cảnh. Nói cách khác, V. và K là các cách biểu diễn khác nhau của cùng một dữ liệu trong cơ chế chú ý cặp khóa-giá trị của chúng. Đặc biệt, Q; K và V là ba các cách trình bày khác nhau của cùng một dữ liệu trong quá trình tự chú ý cơ chế [16].

Khi trọng số và giá trị chú ý được tính toán, bối cảnh vectơ c được tính bằng:

c ¼ /ðfaig; fvigÞ; ð8Þ

trong đó / là hàm trả về một vectơ đơn cho tập hợp giá trị và trọng số tương ứng của chúng.

Cách thực hiện phổ biến của hàm / là thực hiện một tổng trọng số của V:

zi ¼ aivi; ð9Þ

Và

$$c = Xn_{\text{tôi¼1}}^{\text{tử;}} \quad \text{ð10Þ}$$

trong đó zi là biểu diễn có trọng số của một phần tử theo các giá trị và n là thứ nguyên của Z. Ngoài ra còn có một cách khác để thực hiện hàm /, sẽ được trình bày chi tiết ở Phần 3.1. Hoặc theo cách này, vectơ ngữ cảnh sẽ được xác định chủ yếu do trọng số chú ý cao hơn liên quan đến giá trị.

Trên đây là mô tả của chúng tôi về các kiến trúc phổ biến trong mô hình chú ý. Ở đây chúng tôi trích dẫn từ Vaswani et al. [16], cơ chế chú ý ''có thể được mô tả như ánh xạ một truy vấn và một tập hợp các cặp khóa-giá trị thành đầu ra, trong đó truy vấn, khóa, giá trị và đầu ra đều là vectơ . Đầu ra được tính toán dưới dạng trọng số tổng các giá trị, trong đó trọng số được gán cho mỗi giá trị được tính bằng hàm tương thích của truy vấn với khóa tương ứng.”.

Ngoài ra, chúng tôi muốn thảo luận về chỉ số hiệu suất để đánh giá cơ chế chú ý. Nhìn chung, các cơ chế chú ý trong học sâu được gắn với các mô hình mạng lưới thần kinh để nâng cao khả năng xử lý thông tin của họ. Vì thế, thật khó để đánh giá hiệu suất của cơ chế chú ý mà không cần nghiên cứu sâu các mô hình học tập. Một cách tiếp cận phổ biến là nghiên cứu cắt bỏ có nghĩa là để phân tích khoảng cách hiệu suất giữa các mô hình có/không có cơ chế chú ý Ngoài ra, cơ chế chú ý có thể được đánh giá bằng cách hình dung mức độ chú ý (như thể hiện trong Hình 2), nhưng cách này không thể định lượng được.

3. Phân loại sự chú ý

Trong phần trước, chúng tôi đã tóm tắt sự chú ý chung mô hình và giải thích chi tiết từng bước thực hiện cơ chế chú ý. Là một phương pháp để cải thiện việc xử lý thông tin khả năng của mạng lưới thần kinh, cơ chế chú ý có thể được áp dụng cho hầu hết các mô hình trong các lĩnh vực học sâu khác nhau. Mặc dù nguyên tắc của các mô hình chú ý là như nhau nhưng các nhà nghiên cứu đã đưa ra một số sửa đổi và cải tiến cơ chế chú ý trong để thích ứng tốt hơn với các nhiệm vụ cụ thể. Chúng tôi phân loại các cơ chế chú ý theo bốn tiêu chí (như trong Bảng 2). Trong phần này, chúng tôi trình bày chi tiết về các loại cơ chế chú ý khác nhau trong từng tiêu chí thông qua việc xem xét một số bài báo chuyên đề. Ngoài ra, chúng tôi muốn nhấn mạnh rằng các cơ chế chú ý ở các tiêu chí khác nhau không loại trừ lẫn nhau, vì vậy có có thể là sự kết hợp của nhiều tiêu chí trong một mô hình chú ý (xem Phần 5 và Bảng 3).

3.1. Sự nhẹ nhàng của sự chú ý

Sự chú ý được đề xuất bởi Bahdanau et al. [8] như đã đề cập ở trên thuộc về sự chú ý nhẹ nhàng (xác định), sử dụng một trung bình có trọng số của tất cả các khóa để xây dựng vectơ ngữ cảnh. Vì sự chú ý nhẹ nhàng, mô-đun chú ý có thể phân biệt được với tôn trọng đầu vào, do đó toàn bộ hệ thống vẫn có thể được đào tạo bởi các phương pháp lan truyền ngược tiêu chuẩn.

Tương ứng, sự chú ý cứng (ngẫu nhiên) đã được đề xuất bởi Xu và cộng sự. [9], trong đó vectơ ngữ cảnh được tính toán từ các khóa được lấy mẫu ngẫu nhiên. Theo cách này, hàm / trong biểu thức. (số 8) có thể được thực hiện như những điều sau đây:

a~ Multinoulliðfaigþ; ð11Þ

Và

ban 2

Bốn tiêu chí để phân loại cơ chế chú ý và các loại chú ý trong mỗi tiêu chí tiêu chuẩn.

Tiêu chuẩn	Kiểu
Sự chú ý nhẹ nhàng	Aoft/cứng, toàn cầu/địa phương
Các hình thức tính năng đầu vào	Theo mặt hàng, theo vị trí
Biểu diễn đầu vào	Khác biệt, tự thân, đồng sự chú ý, phân cấp
Biểu diễn đầu ra	Đầu ra đơn, nhiều đầu, đa chiều

bàn số 3
Ví dụ về sự kết hợp giữa các loại khác nhau.

Thẩm quyền giải quyết	Ứng dụng	Loại			
		Sự chú ý nhẹ nhàng	Các hình thức tính năng đầu vào	Biểu diễn đầu vào	Biểu diễn đầu ra
[8]	Dịch máy	Mềm mại	Mục khôn ngoan	Đặc sắc	Đầu ra đơn
[7]	Phân loại hình ảnh	Cứng	Vị trí khôn ngoan	Đặc sắc	Đầu ra đơn
[16]	Dịch máy	Mềm mại	Mục khôn ngoan	Đặc sắc	Nhiều đầu
[75]	Trả lời câu hỏi trực quan	Mềm mại	Mục khôn ngoan	Đồng chú ý & phân cấp	Đầu ra đơn
[91]	Hiểu ngôn ngữ	Mềm mại	Mục khôn ngoan	Đặc sắc	Đa chiều
[73]	Phân loại hình ảnh	Mềm mại	Vị trí khôn ngoan	Đặc sắc	Đầu ra đơn
[63]	Phân loại tài liệu	Mềm mại	Mục khôn ngoan	Thứ bậc	Đầu ra đơn

$$c = \sum_{i=1}^{L} \alpha_i v_i \qquad (12)$$

So với mô hình chú ý mềm, mô hình chú ý cứng là về mặt tính toán ít tốn kém hơn vì nó không cần tính toán trọng số chú ý của tất cả các phần tử tại mỗi thời điểm. Tuy nhiên, việc đưa ra quyết định khó khăn ở mỗi vị trí của đặc điểm đầu vào sẽ khiến mô-đun không thể phân biệt và khó tối ưu hóa, vì vậy toàn bộ hệ thống có thể được huấn luyện bằng cách tối đa hóa giới hạn dưới biến thiên gần đúng hoặc tương đương bằng REINFORCE [70].

Trên cơ sở này, Lương et al. [14] thu hút sự chú ý toàn cầu và cơ chế chú ý cục bộ cho dịch máy. Toàn cầu sự chú ý tương tự như sự chú ý nhẹ nhàng. Sự chú ý của địa phương có thể được xem như một sự kết hợp thú vị giữa sự chú ý cứng rắn và mềm mỏng, trong đó chỉ một tập hợp con các từ nguồn được xem xét ở một mức độ nào đó. thời gian. Cách tiếp cận này ít tốn kém về mặt tính toán hơn so với toàn cầu chú ý hoặc chú ý nhẹ nhàng. Đồng thời, không giống như sự chú ý chăm chú, Cách tiếp cận này có thể được phân biệt ở hầu hết mọi nơi, giúp việc này trở nên dễ dàng hơn để thực hiện và đào tạo.

3.2. Các dạng tính năng đầu vào

Các cơ chế chú ý có thể được chia thành từng mục và vị trí khôn ngoan tùy theo tính năng đầu vào có phải là một chuỗi hay không của các mặt hàng đó. Sự chú ý theo từng mục yêu cầu đầu vào là các mục rõ ràng hoặc một bước tiền xử lý bổ sung được thêm vào để tạo ra một chuỗi các mục từ dữ liệu nguồn. Ví dụ, mục có thể là một từ được nhúng trong RNNsearch [8] hoặc một bản đồ đặc trưng trong SENet [71]. Trong mô hình chú ý, bộ mã hóa mã hóa từng mục dưới dạng một mã riêng biệt và gán các trọng số khác nhau với chúng trong quá trình giải mã. Ngược lại, sự chú ý về vị trí nhằm vào các nhiệm vụ khó có được các mục đầu vào riêng biệt, và nói chung cơ chế chú ý như vậy được sử dụng trong các nhiệm vụ trực quan. Ví dụ: bộ giải mã xử lý phần cắt đa độ phân giải của hình ảnh đầu vào ở mỗi bước [7,72] hoặc chuyển đổi vùng liên quan đến nhiệm vụ thành một tư thế chuẩn, được mong đợi để đơn giản hóa việc suy luận ở các lớp tiếp theo [73].

Một điểm khác biệt nữa là cách tính toán khi kết hợp với cơ chế chú ý mềm/cứng. Sự chú ý nhẹ nhàng về mặt hàng tính toán trọng số cho từng mục rồi tạo tổ hợp tuyến tính của chúng. Sự chú ý mềm mại theo vị trí chấp nhận toàn bộ bản đồ đặc trưng làm đầu vào và tạo ra một phiên bản được chuyển đổi thông qua mô-đun chú ý. Thay vì kết hợp tuyến tính tất cả các mục, sự chú ý kỹ càng đến từng mục một cách ngẫu nhiên chọn một hoặc một số các mục dựa trên xác suất của chúng. Sự chú ý kỹ lưỡng về vị trí sẽ ngẫu nhiên chọn một tiểu vùng làm đầu vào và vị trí của tiểu vùng được chọn sẽ được mô-đun chú ý tính toán.

3.3. Biểu diễn đầu vào

Có hai đặc điểm về biểu diễn đầu vào trong hầu hết các các mô hình chú ý nêu trên: 1) Các mô hình này bao gồm một đầu vào duy nhất và một chuỗi đầu ra tương ứng; 2) Các phím và các truy vấn thuộc về hai chuỗi độc lập. Trường hợp này sự chú ý được gọi là sự chú ý đặc biệt [74]. Ngoài ra, cơ chế chú ý còn có nhiều dạng biểu diễn đầu vào khác nhau (Hình 5).

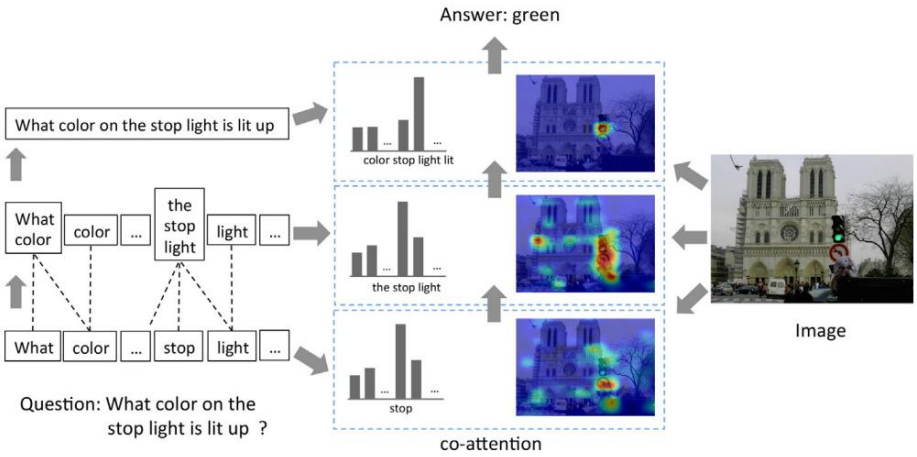
Lu và cộng sự. [75] đã trình bày một mô hình chú ý nhiều đầu vào cho nhiệm vụ trả lời câu hỏi bằng hình ảnh, sự đồng chú ý, cùng lý giải về sự chú ý vào hình ảnh và câu hỏi. Sự đồng chú ý có thể được thực hiện song song hoặc luân phiên. Cái trước tạo ra hình ảnh và câu hỏi chú ý đồng thời, trong khi sau đó tuần tự xen kẽ giữa việc tạo sự chú ý bằng hình ảnh và câu hỏi. Hơn nữa, sự đồng chú ý có thể ở dạng thô hoặc dạng hạt mịn [76]. Sự chú ý chi tiết thô tính toán sự chú ý trên từng đầu vào, sử dụng việc nhúng đầu vào khác dưới dạng truy vấn. Sự chú ý tinh tế đánh giá xem mỗi phần tử của đầu vào ảnh hưởng như thế nào đến từng phần tử của đầu vào khác. Sự đồng chú ý đã được sử dụng thành công trong nhiều lĩnh vực khác nhau. các nhiệm vụ bao gồm phân loại tình cảm [77], khớp văn bản [78], được đặt tên là nhận dạng thực thể [79], định hướng thực thể [80], phân tích nguyên nhân cảm xúc [81] và phân loại tình cảm [82].

Vương và cộng sự. [83] trình bày sự chú ý của bản thân (bên trong) tính toán chỉ chú ý dựa trên trình tự đầu vào. Nói cách khác, sự truy vấn, khóa và giá trị là các cách biểu diễn khác nhau của cùng một trình tự đầu vào. Một mô hình như vậy đã được chứng minh là có hiệu quả bởi một số các tác giả đã khai thác nó theo nhiều cách khác nhau [16,30,84–87]. Một ứng dụng nổi tiếng là Transformer [16], trình tự đầu tiên mô hình tái nạp hoàn toàn dựa trên sự tự chú ý mà không có RNN. Các ứng dụng của mô hình này trong các lĩnh vực khác nhau sẽ được mô tả trong Phần 5.

Trọng số chú ý có thể được tính toán không chỉ từ bản gốc trình tự đầu vào mà còn từ các mức độ trừu tượng khác nhau, chúng tôi gọi là sự chú ý theo thứ bậc. Yang và cộng sự. [63] đề xuất một mạng chú ý phân cấp (HAM) để phân loại tài liệu, có hai cấp độ cơ chế chú ý: cấp độ từ và cấp độ câu. Sự chú ý theo thứ bậc cho phép HAM tổng hợp các từ quan trọng thành một câu rồi tổng hợp câu quan trọng cho một tài liệu. Hơn nữa, hệ thống phân cấp có thể được mở rộng hơn nữa. Wu và cộng sự. [88] đã thêm cấp độ người dùng lên trên cùng, cũng áp dụng sự chú ý ở cấp độ tài liệu. Trái ngược với mô hình học tập trọng lượng trên từ cấp độ thấp hơn đến cấp độ cao hơn, mô hình do Zhao và Zhang đề xuất [61] cũng sử dụng mô hình phân cấp sự chú ý nhưng trọng lượng chú ý được học từ cấp độ cao hơn xuống mức thấp hơn. Ngoại trừ xử lý ngôn ngữ tự nhiên, phân cấp sự chú ý cũng được sử dụng trong thị giác máy tính. Ví dụ, Xiao và cộng sự. [89] đề xuất một phương pháp chú ý hai cấp độ, đó là mức độ chú ý đối tượng và mức độ bộ phận. Đây là phương pháp phân loại hình ảnh chi tiết đầu tiên không sử dụng thông tin bộ phận bổ sung và chỉ dựa vào mô hình để tạo ra trọng số chú ý.

3.4. Biểu diễn đầu ra

Trong phần này, chúng ta thảo luận về các loại biểu diễn đầu ra khác nhau trong các mô hình chú ý Trong số đó, cái phổ biến là đầu ra đơn sự chú ý đề cập đến một biểu diễn tính năng duy nhất trong mỗi lần bước chân. Cụ thể, điểm năng lượng được biểu thị bằng một và chỉ



Hình 5. Một ví dụ về sự chú ý và đồng chú ý theo thứ bậc. Hình từ [75].

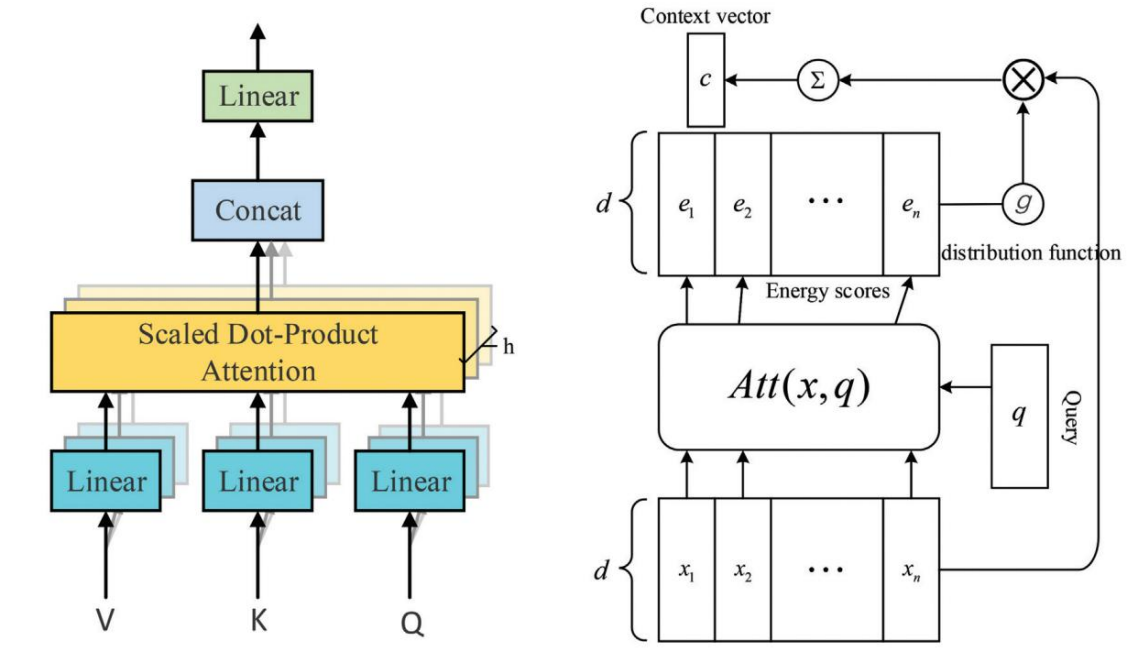
một vectơ tại mỗi bước thời gian. Tuy nhiên, trong một số trường hợp, việc sử dụng một biểu diễn đối tượng duy nhất có thể không thể hoàn thành tốt các nhiệm vụ tiếp theo. Tiếp theo, chúng tôi mô tả hai mô hình chú ý đa đầu ra: đa đầu và đa chiều như minh họa trong Hình 6.

Trong nhiều ứng dụng của mạng nơ ron tích chập, người ta đã chứng minh rằng nhiều kênh có thể thể hiện dữ liệu đầu vào một cách toàn diện hơn so với một kênh. Ngoài ra, trong các mô hình chú ý, trong một số trường hợp, việc sử dụng phân phối chú ý duy nhất của chuỗi đầu vào có thể không đủ cho các tác vụ xuôi dòng. Vaswani và cộng sự. [16] đề xuất sự chú ý nhiều đầu chiếu tuyến tính chuỗi đầu vào (Q; K; V) tới nhiều không gian con dựa trên các tham số có thể học được, sau đó áp dụng sự chú ý tích số chấm theo tỷ lệ cho biểu diễn của nó trong mỗi không gian con và cuối cùng ghép nối đầu ra của chúng .

Bằng cách này, nó cho phép mô hình cùng tham gia vào thông tin từ các không gian con biểu diễn khác nhau ở các vị trí khác nhau. Lý

et al. [90] đã đề xuất ba chính quy hóa bất đồng để tăng cường mô hình chú ý nhiều đầu trên không gian con, vị trí tham dự và biểu diễn đầu ra tương ứng. Cách tiếp cận này khuyến khích sự đa dạng giữa những người đứng đầu chú ý để những người đứng đầu khác nhau có thể tìm hiểu các tính năng riêng biệt và tính hiệu quả được xác nhận thông qua các nhiệm vụ dịch thuật.

Một cách tiếp cận khác là phương pháp đa chiều được đề xuất bởi Shen et al. [91], tính toán vectơ điểm thông minh về tính năng cho các khóa bằng cách thay thế vectơ điểm trọng số bằng ma trận. Bằng cách này, mạng lưới thần kinh có thể tính toán nhiều mức phân bổ sự chú ý cho cùng một dữ liệu. Điều này đặc biệt hữu ích cho quá trình xử lý ngôn ngữ tự nhiên trong đó việc nhúng từ gặp phải vấn đề đa nghĩa. Hơn nữa, Lin và cộng sự. [30] và Du và cộng sự. [92] đã thêm các hình phạt của Frobenius để thực thi sự phân biệt giữa các mô hình liên quan và áp dụng thành công nó vào các nhiệm vụ khác nhau, bao gồm



(a) Multi-head attention [16] (b) Multi-dimensional attention [92]

Hình 6. Minh họa về biểu diễn nhiều đầu ra.

lập hồ sơ tác giả, phân loại tình cảm, sự kế thừa văn bản, và
trích xuất quan hệ được giám sát từ xa.

4. Kiến trúc mạng chú ý

Trong phần này, chúng tôi mô tả ba kiến trúc mạng nơ-ron
được sử dụng kết hợp với sự chú ý. Đầu tiên chúng tôi giải thích chi tiết về
khung mã hóa-giải mã được sử dụng bởi hầu hết các mô hình chú ý. Sau đó
chúng tôi mô tả kiến trúc mạng đặc biệt của các mô hình chú ý,
mạng bộ nhớ, được trang bị bộ nhớ ngoài dài hạn
ký ức. Cuối cùng, chúng tôi giới thiệu các cấu trúc mạng lưới thần kinh đặc biệt
kết hợp với cơ chế chú ý, trong đó RNN không
được sử dụng trong quá trình nắm bắt các phụ thuộc đường dài.

4.1. Bộ mã hóa-giải mã

Bộ mã hóa-giải mã (như trong Hình 7) là một khung chung
dựa trên mạng lưới thần kinh, nhằm mục đích xử lý việc ánh xạ
giữa đầu vào và đầu ra có cấu trúc cao. Ở đây, trước tiên chúng tôi tóm tắt ngắn gọn
mô tả khung mã hóa-giải mã RNN do Cho đề xuất
et al. [5] và Sutskever và cộng sự. [13] cho các nhiệm vụ dịch máy.
Kiến trúc này bao gồm hai RNN hoạt động như một bộ mã hóa và
một cặp giải mã.

Bộ mã hóa ánh xạ chuỗi nguồn có độ dài thay đổi
x ¼ ðx1; ; xTx Þ thành vectơ có độ dài cố định c. Phổ biến nhất
Cách tiếp cận là sử dụng RNN sao cho

$$ht ¼ f xt ð; ht1 Þ ; \quad ð13Þ$$

Và

$$ð c ¼ qf h1; \quad ; hTx g Þ; \quad ð14Þ$$

trong đó ht là trạng thái ẩn tại bước thời gian t và c là vectơ ngữ cảnh
được tạo ra từ chuỗi các trạng thái ẩn. f và q là
chức năng kích hoạt phi tuyến tính. Ví dụ, f có thể đơn giản như
như một hàm sigmoid logistic theo phần tử và phức tạp như một
GRU [5] hoặc LSTM [4], q có thể là trạng thái ẩn hTx của lần cuối
bước thời gian.

Bộ giải mã được huấn luyện để tạo ra mục tiêu có độ dài thay đổi
dãy y ¼ ðy1; ; yTy Þ bằng cách dự đoán ký hiệu tiếp theo yt đã cho
trạng thái ẩn ht. Tuy nhiên, không giống như bộ mã hóa, cả yt và ht
cũng được điều chỉnh theo yt1 và vectơ ngữ cảnh c. Vì thế
trạng thái ẩn của bộ giải mã ở mỗi bước t được tính bằng:

$$ht ¼ fðht1; yt1; cÞ: \quad ð15Þ$$

Mỗi xác suất có điều kiện được mô hình hóa như

$$pðytjfy1; \quad ; yt1g; xÞ ¼ gðyt1; ht; cÞ: \quad ð16Þ$$

Bộ mã hóa và bộ giải mã của mô hình đề xuất được kết hợp với nhau
được huấn luyện để tối đa hóa xác suất có điều kiện của mục tiêu
trình tự cho trước một trình tự nguồn. Một vấn đề tiềm ẩn với
khuôn khổ là mạng lưới thần kinh cần có khả năng nén
tất cả thông tin cần thiết của dữ liệu nguồn thành một đoạn có độ dài cố định
vectơ . Điều này có thể gây khó khăn cho mạng lưới thần kinh để đối phó
với những câu dài, đặc biệt là những câu dài hơn các câu trong ngữ liệu huấn
luyện. Vấn đề này có thể được giảm bớt bằng cách
thêm cơ chế chú ý vào khung bộ mã hóa-giải mã:

$$ð ct ¼ att fh1; \quad ; hTx g Þ; \quad ð17Þ$$

Và

$$pðytjfy1; \quad ; yt1g; xÞ ¼ gðyt1; ht; ctÞ: \quad ð18Þ$$

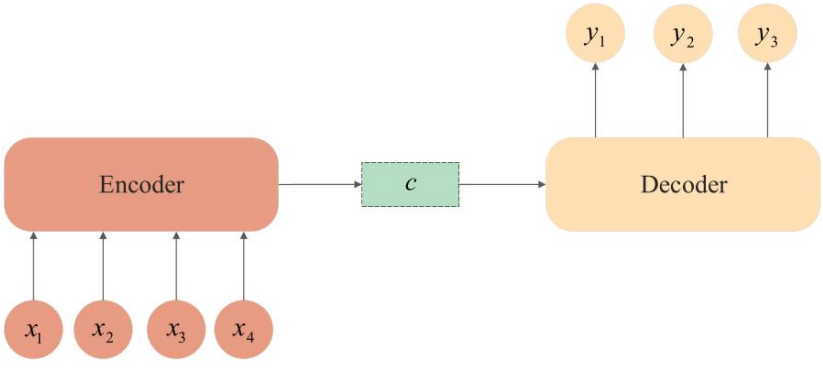
trong đó att là phương thức chú ý được đề cập ở Phần 2 và ct là
vectơ ngữ cảnh được tạo ra bởi cơ chế chú ý khi
giải mã tại thời điểm t. Sự ra đời của cơ chế chú ý
có thể đảm bảo rằng sự đóng góp của các phần tử trong nguồn
trình tự khác nhau khi giải mã các phần tử mục tiêu khác nhau,
như thể hiện trong Hình 8.

Vì khung mã hóa-giải mã này không giới hạn độ dài
của các chuỗi đầu vào và đầu ra, nó có nhiều ứng dụng, chẳng hạn như chú thích
hình ảnh và video [9,93,94], hệ thống hộp thoại tổng quát [95], trả lời câu hỏi
trực quan [75] và lời nói
công nhận [24,96]. Hơn nữa, bộ mã hóa và bộ giải mã cũng có thể
được xây dựng bằng các kiến trúc khác, không nhất thiết chỉ có RNN.
Mặc dù mô hình chú ý có thể được coi là một ý tưởng chung
và không phụ thuộc vào một khuôn khổ cụ thể [7,97,73], hầu hết
các mô hình chú ý hiện được đi kèm với khung bộ mã hóa-giải mã.

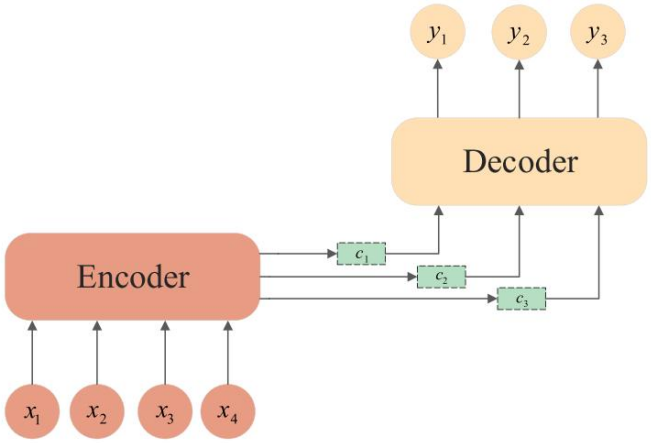
4.2. Mạng bộ nhớ

Ngoài khung mã hóa-giải mã ở phần trước
phần này, cơ chế chú ý cũng được sử dụng kết hợp với
mạng bộ nhớ. Lấy cảm hứng từ cơ chế hoạt động của bộ não con người
xử lý tình trạng quá tải thông tin, mạng bộ nhớ sẽ giới thiệu
bộ nhớ ngoài bổ sung vào mạng lưới thần kinh. Cụ thể, các mạng bộ nhớ
[99,60,98,100,101] lưu một số thông tin liên quan đến tác vụ trong bộ nhớ phụ
bằng cách đưa vào các phần phụ trợ bên ngoài.
đơn vị bộ nhớ và sau đó đọc nó khi cần thiết, điều này không chỉ
tăng hiệu quả dung lượng mạng mà còn cải thiện
hiệu quả tính toán mạng So với sự quan tâm chung
cơ chế, mạng bộ nhớ thay thế khóa bằng bộ nhớ phụ dài hạn và sau đó khớp nội
dung thông qua
cơ chế chú ý

Một ví dụ nổi tiếng là bộ nhớ đầu cuối có khả năng phân biệt
mạng được đề xuất bởi [98], có thể đọc thông tin từ bên ngoài



Hình 7. Minh họa khung mã hóa-giải mã không có cơ chế chú ý.



Hình 8. Minh họa khung mã hóa-giải mã bằng cách sử dụng chú ý cơ chế.

thông tin cuối cùng nhiều lần. Ý tư ởng cốt lõi là chuyển đổi đầu vào đư ợc đặt thành hai đơ n vị bộ nhớ ngoài, một đơ n vị để đánh địa chỉ, và một cái khác cho đầu ra, như trong Hình 9.10. Bộ nhớ đầu cuối mạng có thể đư ợc coi là một dạng của sự chú ý: cặp khóa-giá trị cơ chế chú ý Khác với sự chú ý thông thư ờng, thay vì chỉ mô hình hóa sự chú ý trên một chuỗi duy nhất, họ sử dụng hai đơ n vị bộ nhớ bên ngoài để mô hình hóa nó trên một cơ sở dữ liệu lớn về các chuỗi. TRONG nói cách khác, chúng ta có thể coi cơ chế chú ý như một giao diện tách biệt việc lư u trữ thông tin khỏi việc tính toán, để dung lư ợng mạng có thể đư ợc tăng lên đáng kể chỉ với một lư ợng nhỏ tăng tham số mạng.

4.3. Mạng không có RNN

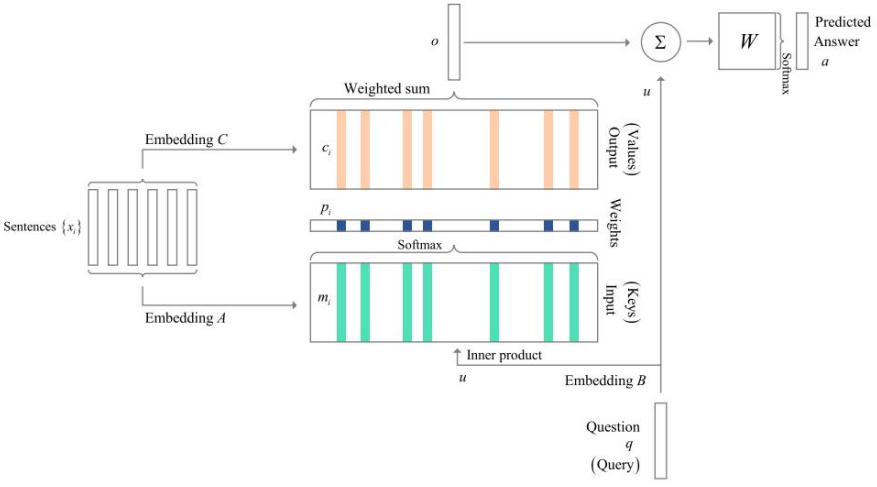
Như đã đề cập ở trên, cả bộ mã hóa và bộ giải mã trong khung mã hóa-giải mã có thể đư ợc triển khai theo nhiều cách như trong Hình 10. RNN dựa trên kiến trúc bộ mã hóa-giải mã thư ờng tính toán nhân tử cùng với vị trí ký hiệu của trình tự đầu vào và đầu ra. Bản chất tuần tự vốn có này dẫn đến tính toán kém hiệu quả vì quá trình xử lý không thể song song. Mặt khác, việc nắm bắt các đối tư ợng phụ thuộc ở khoảng cách xa là điều cần thiết vì cơ chế chú ý trong bộ mã hóa-Khung giải mã cần thu đư ợc thông tin theo ngữ cảnh. Tuy nhiên, độ phức tạp tính toán của việc thiết lập khoảng cách xa sự phụ thuộc của chuỗi có độ dài n qua RNN là O(n). Trong này phần này, chúng tôi mô tả các cách triển khai khác của bộ mã hóa-giải mã khung kết hợp với cơ chế chú ý, loại bỏ thẻ RNN.

Gehring và cộng sự. [102] đề xuất kiến trúc bộ mã hóa-giải mã hoàn toàn dựa vào mạng lư ới thần kinh tích chập kết hợp với cơ chế chú ý. Ngư ợc lại với thực tế là các mạng lặp lại duy trì trạng thái ẩn của toàn bộ quá khứ, các mạng tích chập không dựa vào các tính toán của quá khứ trư ớc đó. bư ớc thời gian, sao cho nó cho phép thực hiện song song trên từng phần tử trong một sự liên tiếp. Kiến trúc này cho phép mạng nắm bắt đư ợc sự phụ thuộc ở khoảng cách xa bằng cách xếp chồng nhiều lớp CNN, độ phức tạp tính toán trở thành O(n=k) đối với CNN nhiều lớp với kích thư ớc hạt tích chập là k. Hơ n nữa, sự tích chập này phư ơ ng pháp có thể khám phá cấu trúc thành phần trong trình tự dễ dàng hơ n vì sự biểu diễn có thứ bậc của nó.

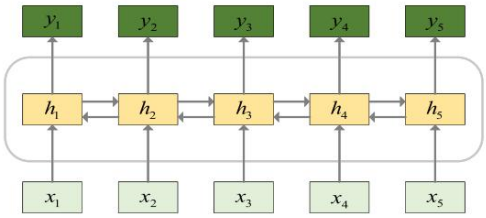
Vaswani và cộng sự. [16] đề xuất một kiến trúc mạng khác, Máy biến áp hoạt động hoàn toàn dựa vào cơ chế tự chú ý để tính toán các biểu diễn đầu vào và đầu ra của nó mà không cần dùng đến RNN hoặc CNN. Transformer bao gồm hai thành phần: lớp mạng chuyển tiếp nguồn cấp dữ liệu theo vị trí (FFN) và lớp chú ý nhiều đầu. FFN theo vị trí là mạng chuyển tiếp nguồn cấp dữ liệu đư ợc kết nối đầy đủ, đư ợc áp dụng riêng cho từng vị trí và giống hệt nhau. Phư ơ ng pháp này có thể đảm bảo thông tin vị trí của từng ký hiệu trong chuỗi đầu vào trong quá trình hoạt động. Sự chú ý của nhiều đầu cho phép mô hình tập trung vào thông tin từ các không gian con biểu diễn khác nhau từ các vị trí khác nhau bằng cách xếp chồng nhiều lớp tự chú ý, giống như nhiều kênh của CNN. Ngoài khả năng song song hóa hơ n, tính phức tạp của việc thiết lập sự phụ thuộc đư ờng dài thông qua cơ chế tự chú ý là O(1).

5. Ứng dụng

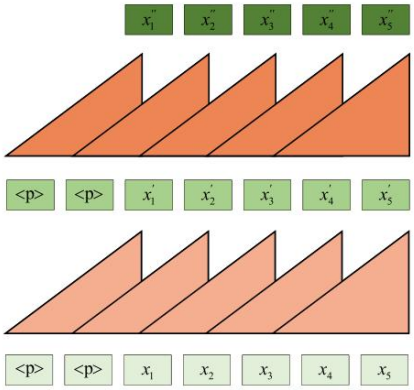
Trong các phần trư ớc, chúng tôi đã chỉ ra rằng các mô hình chú ý có đư ợc ứng dụng rộng rãi vào nhiều công việc khác nhau. Sau đây chúng tôi giới thiệu một số ứng dụng cụ thể của mô hình chú ý trong CV và NLP. chúng tôi làm không có ý định cung cấp một đánh giá toàn diện về tất cả các hệ thống thần kinh kiến trúc sử dụng cơ chế chú ý như ng tập trung vào một số phư ơ ng pháp chú ý chung trong các mô hình chú ý.



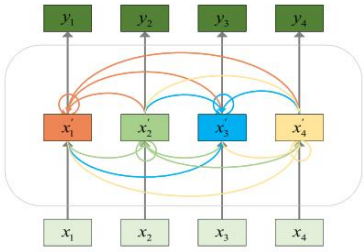
Hình 9. Phiên bản một lớp của mạng bộ nhớ đầu cuối [98]. Ở đây, câu hỏi, đầu vào và đầu ra tư ơ ng ứng với truy vấn, khóa và giá trị trong sự chú ý thống nhất mô hình tư ơ ng ứng.



(a) Single-layer RNN model structure.



(b) Multi-layer CNN model structure.



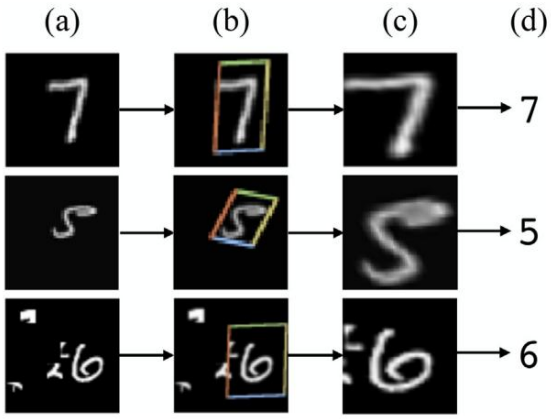
(c) Single-layer self-attention model structure.

Hình 10. Minh họa ba cấu trúc đư ợc sử dụng để nắm bắt sự phụ thuộc khoảng cách xa.

5.1. Ứng dụng trong thị giác máy tính

Trong phần này, chúng tôi mô tả cơ chế chú ý trong CV bằng cách giới thiệu một số bài viết điển hình về các khía cạnh khác nhau của kiến trúc thần kinh.

Sự chú ý không gian cho phép mạng lư ới thần kinh tìm hiểu các vị trí cần tập trung vào, như trong Hình 11. Thông qua cơ chế chú ý này, thông tin không gian trong ảnh gốc đư ợc chuyển sang không gian khác và thông tin chính đư ợc giữ lại. Mnih và cộng sự. [7] đã trình bày một mô hình chú ý không gian đư ợc xây dựng dư ới dạng một RNN duy nhất lấy cửa sổ nhìn thoáng qua làm đầu vào và sử dụng trạng thái bên trong của mạng để chọn vị trí tiếp theo cần tập trung cũng như tạo tín hiệu điều khiển trong môi trư ờng động . Jaderberg và cộng sự. [73] đã giới thiệu một mạng biến áp không gian có thể phân biệt (STN) có thể tìm ra các khu vực cần đư ợc chú ý trong bản đồ đặc trư ng thông qua các phép biến đổi như cắt xén, dịch thuật, xoay, chia tỷ lệ và nghiêng.

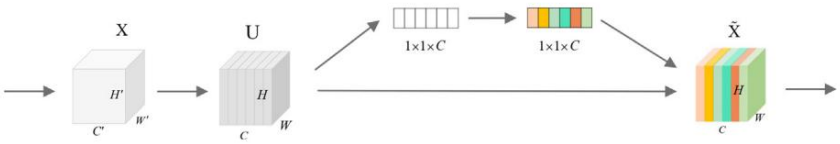


Hình 11. Một ví dụ về sự chú ý không gian từ [73].

Không giống như các lớp gộp, mô-đun biến áp không gian là một cơ chế động có thể chủ động biến đổi không gian một hình ảnh (hoặc bản đồ đặc trư ng) bằng cách tạo ra một phép biến đổi thích hợp cho từng mẫu đầu vào.

Sự chú ý của kênh cho phép mạng lư ới thần kinh tìm hiểu những gì cần tập trung vào, như trong Hình 12. Hu và cộng sự. [71] đã đề xuất mạng ép và kích thích (SE) hiệu chỉnh lại một cách thích ứng các phản hồi tính năng theo kênh bằng cách mô hình hóa rõ ràng sự phụ thuộc lẫn nhau giữa các kênh. Trong mô-đun ép và kích thích đó, nó sử dụng các tính năng tổng hợp trung bình toàn cầu để tính toán mức độ chú ý theo kênh. Li và cộng sự. [103] đã đề xuất SKNet cải thiện hiệu quả và hiệu quả của việc nhận dạng đối tư ợng bằng cách lựa chọn hạt nhân thích ứng theo cách chú ý đến kênh. Ngoài ra, Stollenga và cộng sự. [104] đề xuất một cơ chế chú ý cứng kênh cải thiện hiệu suất phân loại bằng cách cho phép mạng tập trung lặp đi lặp lại vào sự chú ý của các bộ lọc.

Nắm bắt đư ợc sự phụ thuộc tầm xa có tầm quan trọng trung tâm trong mạng lư ới thần kinh sâu, điều này có lợi cho các vấn đề hiểu biết bằng hình ảnh. [87] đã áp dụng cơ chế tự chú ý vào nhiệm vụ thị giác máy tính để giải quyết vấn đề này, đư ợc gọi là sự chú ý không cục bộ, như trong Hình 13. Họ đề xuất mô-đun không cục bộ có mặt nạ chú ý bằng cách tính toán ma trận tư ơ ng quan giữa mỗi điểm không gian trong bản đồ đặc trư ng, sau đó tập trung sự chú ý vào thông tin theo ngữ cảnh dày đặc để tổng hợp. Tuy nhiên, phư ơ ng pháp này cũng có các vấn đề sau: 1) Chỉ liên quan đến mô-đun chú ý vị trí chứ không liên quan đến kênh chú ý thư ờng đư ợc sử dụng



Hình 12. Minh họa kênh chú ý [71].

cơ chế. 2) Khi bản đồ tính năng đầu vào rất lớn thì có vấn đề về hiệu quả thấp. Mặc dù có những phư ơ ng pháp khác để giải quyết vấn đề này, chẳng hạn như chia tỷ lệ, như ng nó sẽ làm mất thông tin và không phải là cách tốt nhất để giải quyết vấn đề này. Để giải quyết những vấn đề này, các nhà nghiên cứu đã cải tiến phư ơ ng pháp phi cục bộ và kết hợp kênh chú ý để đề xuất sự chú ý hỗn hợp [105–111].

Khác với các nghiên cứu trư ớc đây, CCNet [112] đã sử dụng sự chú ý theo vị trí, tạo ra một bản đồ chú ý khổng lồ để ghi lại mối quan hệ giữa từng cặp pixel trong bản đồ đặc trư ng, như trong Hình 14. Mô-đun chú ý chéo chéo col -thông tin theo ngữ cảnh đư ợc chọn lọc theo hư ớng ngang và dọc để nâng cao khả năng đại diện theo pixel. Hơ n nữa, sự chú ý chéo thư ờng xuyên cho phép thu thập thông tin theo ngữ cảnh tầm xa dày đặc từ tất cả các pixel với chi phí tính toán và chi phí bộ nhớ ít hơ n.

5.2. Ứng dụng trong xử lý ngôn ngữ tự nhiên

Trong phần này, trư ớc tiên chúng tôi giới thiệu một số phư ơ ng pháp chú ý đư ợc sử dụng trong các nhiệm vụ khác nhau của NLP và sau đó mô tả một số cách biểu diễn từ đào tạo trư ớc phổ biến đư ợc triển khai với cơ chế chú ý cho các nhiệm vụ NLP.

Dịch máy thần kinh sử dụng mạng thần kinh để dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác. Trong quá trình dịch thuật, việc căn chỉnh các câu trong các ngôn ngữ khác nhau là một vấn đề quan trọng, đặc biệt đối với các câu dài. Bahdanau và cộng sự. [8] đã giới thiệu cơ chế chú ý vào mạng lư ới thần kinh để cải thiện khả năng dịch máy thần kinh bằng cách tập trung có chọn lọc vào các phần của câu nguồn trong quá trình dịch. Sau đó, một số công trình đã đư ợc đề xuất cải tiến, chẳng hạn như sự chú ý của địa phư ơ ng [14], sự chú ý có giám sát [113,114], sự chú ý theo cấp bậc [61] và sự chú ý của bản thân [115,16]. Họ đã sử dụng các kiến trúc chú ý khác nhau để cải thiện sự liên kết của các câu và nâng cao hiệu suất dịch thuật.

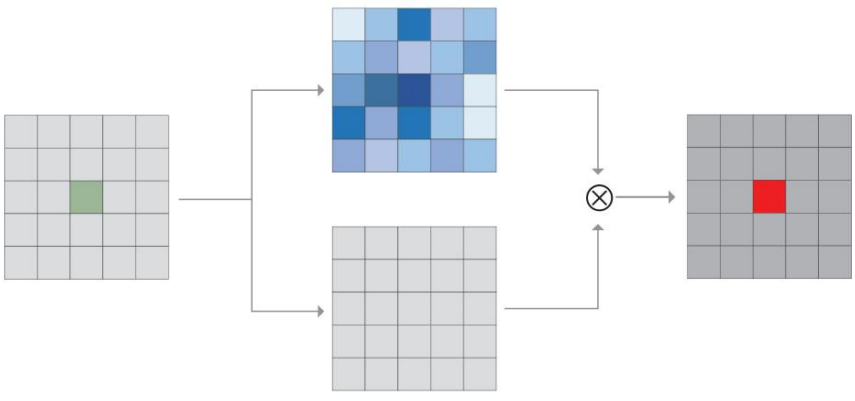
Phân loại văn bản nhằm mục đích gán nhãn cho văn bản và có các ứng dụng rộng rãi bao gồm ghi nhãn chủ đề [116], phân loại tình cảm [117,118] và phát hiện thư rác [119]. Trong các nhiệm vụ phân loại này, sự chú ý của bản thân chủ yếu đư ợc sử dụng để xây dựng cách trình bày tài liệu hiệu quả hơ n [55,91,30]. Trên cơ sở đó, một số công trình đã kết hợp cơ chế tự chú ý với các phư ơ ng pháp chú ý khác, chẳng hạn như tự chú ý theo cấp bậc [63] và tự chú ý đa chiều [30]. Ngoài ra, các tác vụ này còn áp dụng các kiến trúc mô hình chú ý (xem Phần 4) như Transformer [120,121] và mạng bộ nhớ [122,123].

So khớp văn bản cũng là một vấn đề nghiên cứu cốt lõi trong NLP và truy xuất thông tin, bao gồm trả lời câu hỏi, tìm kiếm tài liệu, phân loại kế thừa, nhận dạng diễn giải và đề xuất kèm theo các bài đánh giá. Nhiều nhà nghiên cứu đã đư a ra các phư ơ ng pháp tiếp cận mới kết hợp với khả năng hiểu sự chú ý, chẳng hạn như mạng bộ nhớ [98], chú ý hơ n chú ý [124], chú ý bên trong [83], chú ý có cấu trúc [65] và đồng chú ý [78 ,75].

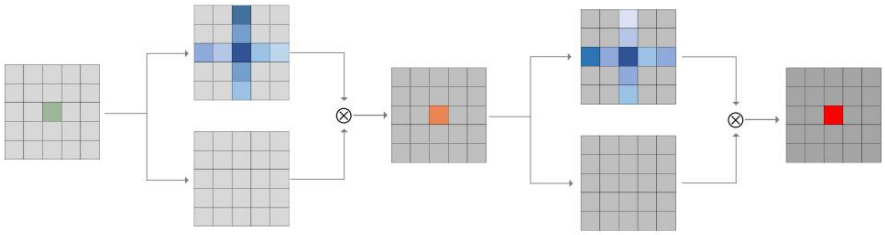
Biểu diễn từ đư ợc đào tạo trư ớc là thành phần chính trong nhiều mô hình hiểu ngôn ngữ thần kinh. Tuy nhiên, các nghiên cứu trư ớc đây [125,126] chỉ xác định đư ợc một cách nhúng cho cùng một từ, điều này không thể đạt đư ợc việc nhúng từ theo ngữ cảnh. Peters và cộng sự. [127] đã giới thiệu một cách tiếp cận chung về biểu diễn phụ thuộc vào ngữ cảnh với Bi-LSTM để giải quyết vấn đề này. Lấy cảm hứng từ mô hình Máy biến áp [62], các nhà nghiên cứu đã đề xuất các biểu diễn bộ mã hóa hai chiều từ máy biến áp (BERT) [128] và phư ơ ng pháp đào tạo trư ớc tổng quát (GPT) [129–131] theo các bộ phận mã hóa và giải mã. BERT là mô hình ngôn ngữ hai chiều và có hai nhiệm vụ đào tạo trư ớc sau: 1) Mô hình ngôn ngữ mặt nạ (MLM). Nó chỉ đơ n giản che giấu một số phần trăm mã thông báo đầu vào một cách ngẫu nhiên và sau đó dự đoán các mã thông báo bị che giấu đó. 2) Dự đoán câu tiếp theo. Nó sử dụng bộ phân loại nhị phân tuyến tính để xác định xem hai câu có đư ợc kết nối hay không. GPT là mô hình một chiều và phư ơ ng pháp đào tạo của nó đại khái là sử dụng từ trư ớc đó để dự đoán từ tiếp theo. Các thử nghiệm cho thấy những cải tiến lớn khi áp dụng chúng vào nhiều nhiệm vụ NLP.

6. Chú ý đến khả năng diễn giải

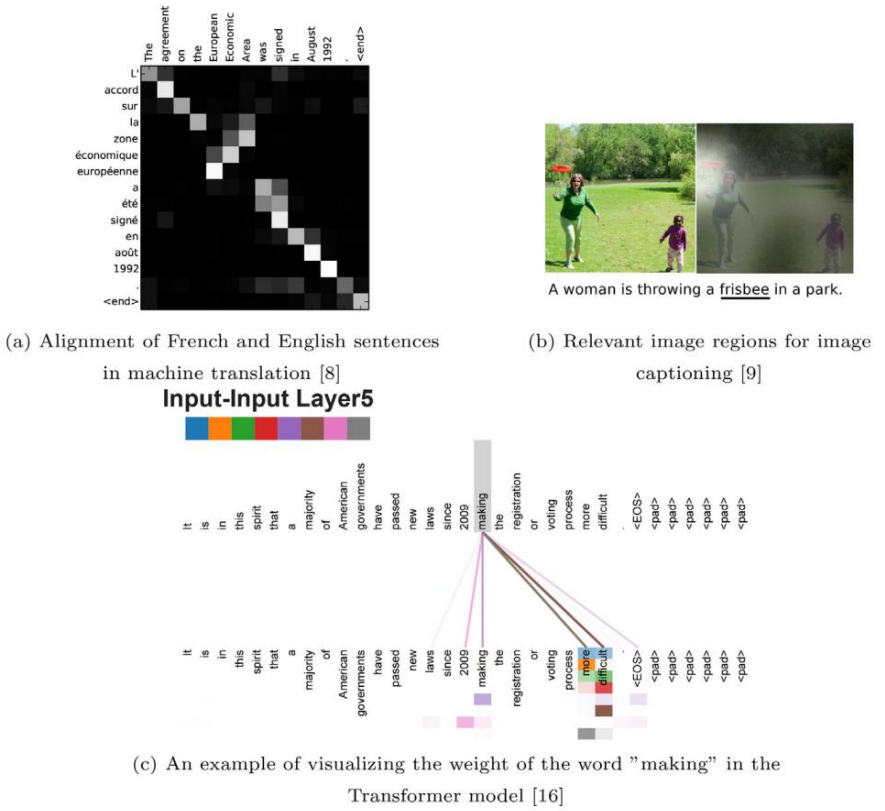
Trong những năm gần đây, trí tuệ nhân tạo đã đư ợc phát triển nhanh chóng [132–137], đặc biệt là trong lĩnh vực học sâu. Làm sao-



Hình 13. Minh họa sự chú ý không cục bộ [87].



Hình 14. Minh họa sự chú ý chéo [112].



(a) Alignment of French and English sentences in machine translation [8]

(b) Relevant image regions for image captioning [9]

(c) An example of visualizing the weight of the word "making" in the Transformer model [16]

Hình 15. Ví dụ về hình dung sức nặng của cơ chế chú ý.

Chư a bao giờ, khả năng diễn giải là mối quan tâm lớn đối với nhiều mô hình học sâu hiện nay. Các mô hình học sâu ngày càng trở nên phức tạp, do đó, điều quan trọng là phải tìm hiểu các chức năng ra quyết định từ dữ liệu và đảm bảo rằng chúng ta hiểu lý do tại sao một quyết định cụ thể lại xảy ra [51]. Lớp chú ý trong mô hình mạng thần kinh cung cấp cách suy luận về mô hình đằng sau các dự đoán của nó [54–57,138], như ng điều này thư ờng bị chỉ trích là không rõ ràng [52,53].

Như đư ợc hiển thị trong Hình 15(a), Bahdanau et al. [8] chú thích trọng số chú ý đư ợc trực quan hóa, hiển thị rõ ràng sự liên kết (mềm) giữa các từ trong bản dịch đư ợc tạo (tiếng Pháp) và các từ trong câu nguồn (tiếng Anh). Hình 15(b) hiển thị các vùng tham dự tư ơ ng ứng với từ đư ợc gạch chân trong quá trình tạo chú thích hình ảnh thần kinh [9]. Như chúng ta có thể thấy, mô hình học cách sắp xếp rất phù hợp với trực giác của con ngư ời.

Hơ n nữa, Hình 15(c) hiển thị một ví dụ về trực quan hóa trong lớp tự chú ý của bộ mã hóa trong mô hình Transformer do Vaswani et al đề xuất. [16]. Màu sắc khác nhau đại diện cho những cái đầu khác nhau.
Hơ n nữa, Voita et al. [139] đã đánh giá sự đóng góp của các cá nhân quan tâm đến mô hình Transformer về dịch thuật. Họ phát hiện ra rằng những cái đầu khác nhau cho thấy mức độ quan trọng khác nhau thông qua các hoạt động cắt tỉa. Ngoài ra, Chan và cộng sự. [24] quan sát thấy cơ chế chú ý dựa trên nội dung có thể xác định chính xác vị trí bắt đầu trong chuỗi âm thanh cho ký tự đầu tiên.

Tuy nhiên, một số nghiên cứu gần đây [52,53] cho rằng sự chú ý không thể đư ợc coi là một phư ơ ng tiện đáng tin cậy để giải thích mạng lư ới thần kinh sâu. Jain và Wallace [52] đã thực hiện các thử nghiệm sâu rộng trên nhiều nhiệm vụ NLP khác nhau và chứng minh rằng sự chú ý không nhất quán với các số liệu có thể giải thích khác. Họ nhận thấy rằng việc xây dựng các phân bổ sự chú ý đối nghịch là rất thư ờng xuyên, điều đó có nghĩa là các phân bổ trọng số chú ý khác nhau mang lại những dự đoán tư ơ ng đư ơ ng. Serrano và Smith [53] đã áp dụng một phân tích khác dựa trên biểu diễn trung gian và nhận thấy rằng trọng số chú ý chỉ là những yếu tố dự báo ồn ào về tầm quan trọng của các thành phần trung gian. Ngư ợc lại, Wiegreffe và Pinter [57] đã bác bỏ công việc của họ bằng bốn thử nghiệm thay thế. Vì vậy, liệu sự chú ý có đư ợc sử dụng như một phư ơ ng tiện để giải thích mạng lư ới thần kinh hay không vẫn là một chủ đề mở.

7. Những thách thức và triển vọng

Các mô hình chú ý đã trở nên phổ biến trong các mạng lư ới thần kinh sâu. Vì Bahdanau et al. [8] đã sử dụng cơ chế chú ý

để căn chỉnh các tác vụ dịch máy, nhiều biến thể cơ chế chú ý khác nhau đã xuất hiện không ngừng. Mô hình Transformer chỉ sử dụng cơ chế tự chú ý do Vaswani et al đề xuất. [16] cũng là một cột mốc quan trọng trong cơ chế chú ý và các biến thể của nó [128,129,140–142] đã đư ợc áp dụng thành công trong nhiều lĩnh vực khác nhau. Khi nghiên cứu mô hình chú ý, chúng tôi thấy rằng cơ chế chú ý có tính đổi mới ở một số khía cạnh, chẳng hạn như triển khai nhiều chức năng điểm số và chức năng phân phối, sự kết hợp giữa giá trị và trọng số chú ý cũng như kiến trúc mạng.

Vẫn còn nhiều chỗ cần cải thiện trong việc thiết kế các cơ chế chú ý. Ở đây chúng tôi tóm tắt một số hư ớng đi đầy hứa hẹn như sau.

Hiện nay, hầu hết các mô hình tự chú ý biểu diễn truy vấn và khóa một cách độc lập, tuy nhiên, một số nghiên cứu gần đây [143,144] đã đạt đư ợc hiệu suất tốt bằng cách kết hợp truy vấn và khóa.
Liệu truy vấn và khóa có cần thiết để tồn tại độc lập trong quá trình tự chú ý hay không vẫn là một câu hỏi mở.
Hàm phân phối chú ý có ảnh hư ởng lớn đến độ phức tạp tính toán của toàn bộ mô hình chú ý. Một số nghiên cứu gần đây [145,146] cho thấy độ phức tạp của việc tính toán sự chú ý có thể giảm hơ n nữa bằng cách cải thiện chức năng phân phối sự chú ý.

Làm thế nào các kỹ thuật chú ý đư ợc phát triển trong một lĩnh vực có thể đư ợc áp dụng cho các lĩnh vực khác cũng là một hư ớng đi thú vị. Ví dụ: khi áp dụng phư ơ ng pháp chú ý với tính năng tự chú ý trong NLP trong CV, nó sẽ cải thiện hiệu suất đồng thời giảm hiệu quả, giống như mạng không cục bộ [87].

Sự kết hợp giữa cơ chế thích ứng và cơ chế chú ý có thể tự động đạt đư ợc hiệu quả của sự chú ý theo cấp bậc mà không cần thiết kế cấu trúc của từng lớp theo cách thủ công.
Khám phá các chỉ số hiệu suất hiệu quả hơ n để đánh giá cơ chế chú ý cũng là một chủ đề thú vị. Sen và cộng sự. [147] đã thiết kế một tập hợp các phư ơ ng pháp đánh giá để định lư ợng sự tư ơ ng đồng giữa mô hình thần kinh dựa trên sự chú ý và con ngư ời bằng cách sử dụng các số liệu tư ơ ng tự về bản đồ chú ý mới. Đây cũng là hư ớng mở cho các nghiên cứu tiếp theo.

8. Kết luận

Trong bài viết này, chúng tôi minh họa cách hoạt động của mô hình chú ý thống nhất và mô tả chi tiết việc phân loại các cơ chế chú ý. Sau đó, chúng tôi tóm tắt các kiến trúc mạng đư ợc sử dụng kết hợp với cơ chế chú ý và giới thiệu các ứng dụng điển hình của cơ chế chú ý đư ợc sử dụng trong thị giác máy tính và xử lý ngôn ngữ tự nhiên. Cuối cùng, chúng tôi thảo luận về khả năng dự đoán lẫn nhau mà các cơ chế chú ý mang lại cho quá trình tạo mô hình, những thách thức và triển vọng của các mô hình chú ý hiện tại. Tóm lại, cơ chế chú ý đã đư ợc sử dụng thành công trong nhiều lĩnh vực ứng dụng học sâu khác nhau, như ng vẫn còn nhiều câu hỏi thú vị cần đư ợc khám phá.

Tuyên bố đóng góp quyền tác giả CRediT

Zhaoyang Niu: Điều tra, Phư ơ ng pháp luận, Phân tích hình thức, Viết - ôn tập & biên tập. Guoqiang Zhong: Khái niệm hóa, Phư ơ ng pháp luận, Điều tra, Phân tích chính thức, Viết - đánh giá & chỉnh sửa, Quản lý dự án, Giám sát. Hui Yu: Khái niệm hóa, Điều tra, Phư ơ ng pháp luận, Viết - đánh giá & chỉnh sửa, Giám sát.

Tuyên bố về lợi ích cạnh tranh

Các tác giả tuyên bố rằng họ không có lợi ích tài chính hoặc mối quan hệ cá nhân cạnh tranh nào có thể ảnh hư ởng đến công việc đư ợc báo cáo trong bài viết này.

Sự nhìn nhận

Công trình này đư ợc hỗ trợ bởi Chư ơ ng trình Nghiên cứu và Phát triển Trọng điểm Quốc gia của Trung Quốc theo Trợ cấp số 1.
2018AAA0100400, Quỹ chung của Nghiên cứu tiền thiết bị và Bộ Giáo dục Trung Quốc theo Khoản tài trợ số 2018.
6141A020337, Quỹ khoa học tự nhiên của tỉnh Sơ n Đông theo số tài trợ ZR2020MF131, và Chư ơ ng trình khoa học và công nghệ của Thanh Đảo theo tài trợ số 21-1-4-ny-19-nsh.

Ngư ời giới thiệu

[1] RA Rensink, Sự thể hiện động của các cảnh, Visual Cogn. 7 (2000) 17–42.
[2] M. Corbetta, GL Shulman, Kiểm soát sự chú ý theo mục tiêu và kích thích trong não, Nat. Mục sư Neurosci. 3 (2002) 201–215.
[3] JK Tsotsos, SM Culhane, WYK Wai, Y. Lai, N. Davis, F. Nuflo, Mô hình hóa sự chú ý trực quan thông qua điều chỉnh có chọn lọc, Artif. Trí tuệ. 78 (1995) 507–545.
[4] S. Hochreiter, J. Schmidhuber, Trí nhớ ngắn hạn dài, Máy tính thần kinh. 9 (1997) 1735–1780.
[5] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Học cách biểu diễn cụm từ bằng bộ mã hóa-giải mã RNN cho dịch máy thống kê, trong: EMNLP, ACL, 2014, trang 1724– 1734..
[6] L. Itti, C. Koch, E. Niebur, Một mô hình chú ý trực quan dựa trên độ mặn để phân tích cảnh nhanh, IEEE Trans. Mẫu Hậu Môn. Mach. Trí tuệ. 20 (1998) 1254–1259.
[7] V. Mnih, N. Heess, A. Graves, K. Kavukcuoglu, Các mô hình chú ý thị giác tái diễn, trong: NIPS, trang 2204–2212..
[8] D. Bahdanau, K. Cho, Y. Bengio, Dịch máy thần kinh bằng cách cùng học cách căn chỉnh và dịch, trong: ICLR..
[9] K. Xu, J. Ba, R. Kiros, K. Cho, AC Courville, R. Salakhutdinov, RS Zemel, Y. Bengio, Show, tham dự và kể: tạo chú thích hình ảnh thần kinh với sự chú ý trực quan, trong: ICML, Tập 37 của Kỷ yếu hội nghị và hội thảo JMLR, JMLR.org, 2015, trang 2048–2057..
[10] J. Lu, C. Xiong, D. Parikh, R. Socher, Biết khi nào cần xem xét: Chú ý thích ứng thông qua trọng điểm trực quan để tạo chú thích cho hình ảnh, trong: CVPR, IEEE Computer Society, 2017, trang 3242–3250.
[11] G. Liu, J. Guo, LSTM hai chiều với cơ chế chú ý và lớp tích chập để phân loại văn bản, Neurocomputing 337 (2019) 325–338.
[12] Y. Li, L. Yang, B. Xu, J. Wang, H. Lin, Cải thiện phân loại thuộc tính ngư ời dùng với sự chú ý của văn bản và mạng xã hội, Cogn. Máy tính. 11 (2019) 459–468.
[13] I. Sutskever, O. Vinyals, QV Le, Học theo trình tự với mạng lư ới thần kinh, trong: NIPS, trang 3104–3112..
[14] T. Luong, H. Pham, CD Manning, Các phư ơ ng pháp tiếp cận hiệu quả đối với dịch máy thần kinh dựa trên sự chú ý, trong: EMNLP, Hiệp hội Ngôn ngữ học tính toán, 2015, trang 1412–1421..
[15] D. Britz, A. Goldie, M. Luong, QV Le, Khám phá lớn về kiến trúc dịch máy thần kinh, CoRR abs/1703.03906 (2017).
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN Gomez, L. Kaiser, I. Polosukhin, Chú ý là tất cả những gì bạn cần, trong: NIPS, trang 5998–6008. .
[17] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, Mô hình chú ý theo không gian-thời gian từ đầu đến cuối để nhận dạng hành động của con ngư ời từ dữ liệu bộ xư ơ ng, trong: AAAI, AAAI Press, 2017 , trang 4263–4270..
[18] Y. Tian, W. Hu, H. Jiang, J. Wu, Mạng dư lư ợng kim tự tháp chú ý đư ợc kết nối dày đặc để ư ớc tính tư thế con ngư ời, Neurocomputing 347 (2019) 13–23.
[19] A. Zhao, L. Qi, J. Li, J. Dong, H. Yu, LSTM để chẩn đoán bệnh thoái hóa thần kinh bằng dữ liệu dáng đi, trong: H. Yu, J. Dong (Eds.), Hội nghị quốc tế lần thứ chín về Xử lý đồ họa và hình ảnh (ICGIP 2017), tập. 10615, tr. 106155B..
[20] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, N. Zheng, Thêm sự chú ý đến các nơ -ron trong mạng lư ới thần kinh tái phát, trong: ECCV (9), Tập 11213 của Ghi chú Bài giảng trong Khoa học Máy tính, Springer, 2018, trang 136–152..
[21] K. Song, T. Yao, Q. Ling, T. Mei, Tăng cư ờng phân tích cảm xúc hình ảnh bằng sự chú ý trực quan, Neurocomputing 312 (2018) 218–228.
[22] X. Yan, S. Hu, Y. Mao, Y. Ye, H. Yu, Phư ơ ng pháp học tập đa góc nhìn sâu: đánh giá, Máy tính thần kinh (2021).
[23] J. Chorowski, D. Bahdanau, K. Cho, Y. Bengio, Nhận dạng giọng nói liên tục từ đầu đến cuối bằng cách sử dụng NN lặp lại dựa trên sự chú ý: kết quả đầu tiên, CoRR abs/1412.1602 (2014)..

[24] W. Chan, N. Jaitly, QV Le, O. Vinyals, Nghe, tham dự và đánh vần: mạng lưới thần kinh để nhận dạng giọng nói đàm thoại từ vựng lớn, trong: ICASSP, IEEE, 2016, trang 4960–4964.

[25] M. Sperber, J. Niehues, G. Neubig, S. Stüker, A. Waibel, Các mô hình âm thanh tự chú ý, trong: INTERSPEECH, ISCA, 2018, trang 3723–3727..

[26] S. Wang, L. Hu, L. Cao, X. Huang, D. Lian, W. Liu, Nhúng bối cảnh giao dịch dựa trên sự chú ý cho đề xuất mục tiếp theo, trong: AAAI, AAAI Press, 2018, tr. 2532–2539.

[27] H. Ying, F. Zhuang, F. Zhang, Y. Liu, G. Xu, X. Xie, H. Xiong, J. Wu, Hệ thống gợi ý tuần tự dựa trên mạng chú ý phân cấp, trong: IJCAI, ijcai. org, 2018, trang 3926–3932..

[28] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Mạng chú ý đồ thị, trong: ICLR (Poster), OpenReview.net, 2018..

[29] K. Xu, L. Wu, Z. Wang, Y. Feng, V. Sheinin, Graph2seq: biểu đồ để học theo trình tự với mạng thần kinh dựa trên sự chú ý, CoRR abs/1804.00823 (2018)..

[30] Z. Lin, M. Feng, CN dos Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, Một câu nhúng có cấu trúc chú ý đến bản thân, trong: ICLR (Poster), OpenReview. mạng, 2017..

[31] K. Zhang, G. Zhong, J. Dong, S. Wang, Y. Wang, Dự đoán thị trường chứng khoán dựa trên mạng lưới đối thủ tổng quát, trong: R. Bie, Y. Sun, J. Yu (Eds.), Hội nghị quốc tế 2018 về nhận dạng, thông tin và kiến thức trong Internet vạn vật, IIKI 2018, Bắc Kinh, Trung Quốc, ngày 19–21 tháng 10 năm 2018, Tập 147 của Khoa học máy tính Procedia, Elsevier, 2018, trang 400–406..

[32] C. Ieracitano, A. Paviglianiti, M. Campolo, A. Hussain, E. Pasero, FC Morabito, Một hệ thống phân loại tự động mới dựa trên phương pháp học máy có giám sát và không giám sát lai cho sợi nano quay điện, IEEE CAA J. Autom. Sinica 8 (2021) 64–76.

[33] Z. Fan, G. Zhong, H. Li, Một mạng tổng hợp tính năng để phát hiện xoáy quy mô trung bình đa phương thức, trong: H. Yang, K. Pasupa, AC Leung, JT Kwok, JH Chan, I. King (Eds.), Xử lý thông tin thần kinh – Hội nghị quốc tế lần thứ 27, ICONIP 2020, Bangkok, Thái Lan, ngày 23–27 tháng 11 năm 2020, Kỷ yếu, Phần I, tập 12532 của Ghi chú Bài giảng về Khoa học Máy tính, Springer, 2020, trang 51–61. .

[34] H. Yu, O. Garrod, R. Jack, P. Schyns, Một khuôn khổ để tạo biểu cảm khuôn mặt tự động và có giá trị về mặt nhận thức, Ứng dụng Công cụ Đa phương tiện. 74 (2015) 9427–9447.

[35] Q. Li, Z. Fan, G. Zhong, Bednet: mạng phát hiện cạnh hai chiều để phát hiện mặt trước đại dương, trong: H. Yang, K. Pasupa, AC Leung, JT Kwok, JH Chan, I. King (Eds.), Xử lý thông tin thần kinh – Hội nghị quốc tế lần thứ 27, ICONIP 2020, Bangkok, Thái Lan, ngày 18–22 tháng 11 năm 2020, Kỷ yếu, Phần IV, tập 1332 của Truyền thông trong Khoa học Thông tin và Máy tính, Springer, 2020, trang 312 –319..

[36] Z. Fan, G. Zhong, H. Wei, H. Li, Ednet: mạng phát hiện xoáy quy mô trung bình với dữ liệu đa phương thức, tại: Hội nghị chung quốc tế về mạng thần kinh năm 2020, IJCNN 2020, Glasgow, Vương quốc Anh, Ngày 19–24 tháng 7 năm 2020, IEEE, 2020, trang 1–7..

[37] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, TD Pham, Mạng lưới thần kinh tích chập phân cấp song song dựa trên khu vực để đánh giá liệt dây thần kinh mặt tự động, IEEE Trans. Hệ thống thần kinh Trại cai nghiện. Anh. 28 (2020) 2325–2332.

[38] W. Yue, Z. Wang, W. Liu, B. Tian, S. Lauria, X. Liu, Một phương pháp lọc cộng tác dựa trên vật phẩm và người dùng có trọng số tối ưu để dự đoán dữ liệu cơ bản cho bệnh nhân mắt điều hòa Friedreich, Neurocomputing 419 (2021) 287–294.

[39] N. Zeng, H. Li, Z. Wang, W. Liu, S. Liu, FE Alsaadi, X. Liu, Phân đoạn hình ảnh dựa trên học tập tăng cường sâu để phân tích định lượng dải sắc ký miễn dịch vàng, Máy tính thần kinh (2020).

[40] W. Liu, Z. Wang, X. Liu, N. Zeng, D. Bell, Một phương pháp tối ưu hóa bầy hạt mới để phân cụm bệnh nhân từ các khoa cấp cứu, IEEE Trans. Tiến hóa. Máy tính. 23 (2019) 632–644.

[41] N. Zeng, Z. Wang, H. Zhang, K.-E. Kim, Y. Li, X. Liu, Một bộ lọc hạt cải tiến với phân phối đề xuất lai mới để phân tích định lượng các dải sắc ký miễn dịch vàng, IEEE Trans. Công nghệ nano. 18 (2019) 819–829 .

[42] Y. Ming, X. Meng, C. Fan, H. Yu, Học sâu để ước tính độ sâu bằng một mắt: đánh giá, Máy tính thần kinh 438 (2021) 14–33.

[43] Y. Xia, H. Yu, F. Wang, Định vị trung tâm mắt chính xác và mạnh mẽ thông qua mạng tích chập hoàn toàn, IEEE CAA J. Autom. Sinica 6 (2019) 1127–1138.

[44] Y. Guo, Y. Xia, J. Wang, H. Yu, R. Chen, Điện toán cảm xúc trên khuôn mặt theo thời gian thực trên thiết bị di động, Cảm biến 20 (2020) 870.

[45] Y. Wang, X. Dong, G. Li, J. Dong, H. Yu, Phân tích mặt trước dựa trên hồi quy Cascade để phân tích biểu cảm khuôn mặt động, Cogn. Máy tính. (2021) 1–14.

[46] X. Zhang, D. Ma, H. Yu, Y. Huang, P. Howell, B. Stevens, Nhận thức cảnh hướng dẫn phát hiện sự bất thường của đám đông, Máy tính thần kinh 414 (2020) 291–302.

[47] A. Roy, B. Banerjee, A. Hussain, S. Poria, Thiết kế từ điển phân biệt để phân loại hành động trong ảnh tĩnh và video, Cogn. Máy tính. (2021).

[48] S. Liu, Y. Xia, Z. Shi, H. Yu, Z. Li, J. Lin, Học sâu trong uốn kim loại tấm với mạng lưới thần kinh sâu hướng dẫn lý thuyết mới, IEEE/CAA J. Autom. Sinica 8 (2021) 565–581.

[49] F. Luque Snchez, I. Hupont, S. Tabik, F. Herrera, Xem lại phân tích hành vi của đám đông thông qua học sâu: phân loại, phát hiện bất thường, cảm xúc của đám đông, bộ dữ liệu, cơ hội và triển vọng, Inf. Hợp nhất 64 (2020) 318– 335.

[50] X. Zhang, X. Yang, W. Zhang, G. Li, H. Yu, Đánh giá cảm xúc đám đông dựa trên suy luận mờ nhạt về kích thích và hóa trị, Neurocomputing 445 (2021) 194–205.

[51] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, Khảo sát các phương pháp giải thích mô hình hộp đen, ACM Comput. Sống sót. 51 (2019) 93:1–93:42..

[52] S. Jain, BC Wallace, Sự chú ý không phải là lời giải thích, trong: NAACL-HLT (1), Hiệp hội Ngôn ngữ học tính toán, 2019, trang 3543–3556..

[53] S. Serrano, NA Smith, Sự chú ý có thể diễn giải được không?, trong: ACL (1), Hiệp hội Ngôn ngữ học tính toán, 2019, trang 2931–2951..

[54] LH Li, M. Yatskar, D. Yin, C. Hsieh, K. Chang, BERT với tầm nhìn nhìn vào cái gì?, trong: ACL, Hiệp hội Ngôn ngữ học tính toán, 2020, trang 5265–5275..

[55] G. Letarte, F. Paradis, P. Giguère, F. Laviolette, Tầm quan trọng của việc tự chú ý đối với phân tích tình cảm, trong: BlackboxNLP@EMNLP, Hiệp hội Ngôn ngữ học tính toán, 2018, trang 267–275..

[56] S. Vashishth, S. Upadhyay, GS Tomar, M. Faruqui, Khả năng diễn giải chú ý qua các nhiệm vụ NLP, CoRR abs/1909.11218 (2019).

[57] S. Wiegreffe, Y. Pinter, Sự chú ý không phải là lời giải thích, trong: EMNLP/IJCNLP (1), Hiệp hội Ngôn ngữ học tính toán, 2019, trang 11–20..

[58] M. Schuster, KK Paliwal, Mạng thần kinh tái phát hai chiều, IEEE Trans. Quá trình tín hiệu. 45 (1997) 2673–2681.

[59] A. Sordoni, P. Bachman, Y. Bengio, Sự chú ý thần kinh xen kẽ lặp đi lặp lại để đọc bằng máy, CoRR abs/1606.02245 (2016)..

[60] A. Graves, G. Wayne, I. Danihelka, Máy turing thần kinh, CoRR abs/1410.5401 (2014).

[61] S. Zhao, Z. Zhang, Dịch máy thần kinh chú ý qua chú ý, trong: AAAI, Nhà xuất bản AAAI, 2018, trang 563–570..

[62] A. Galassi, M. Lippi, P. Torroni, Xin hãy chú ý! Đánh giá quan trọng về các mô hình chú ý thần kinh trong xử lý ngôn ngữ tự nhiên, CoRR abs/1902.02181 (2019).

[63] Z. Yang, D. Yang, C. Dyer, X. He, AJ Smola, EH Hovy, Mạng chú ý phân cấp để phân loại tài liệu, trong: HLT-NAACL, Hiệp hội Ngôn ngữ học tính toán, 2016, trang 1480– 1489..

[64] AFT Martins, RF Astudillo, Từ softmax đến thưa thớt: Một mô hình thưa thớt về sự chú ý và phân loại nhiều nhãn, trong: ICML, Tập 48 của Kỷ yếu Hội thảo và Hội nghị JMLR, JMLR.org, 2016, trang 1614–1623. .

[65] Y. Kim, C. Denton, L. Hoang, AM Rush, Mạng chú ý có cấu trúc, arXiv: Tính toán và Ngôn ngữ (2017)..

[66] AH Miller, A. Fisch, J. Dodge, A. Karimi, A. Bordes, J. Weston, Mạng bộ nhớ khóa-giá trị để đọc trực tiếp tài liệu, trong: EMNLP, Hiệp hội Ngôn ngữ học tính toán, 2016, tr. 1400–1409..

[67] J. Ba, GE Hinton, V. Mnih, JZ Leibo, C. Ionescu, Sử dụng tạ nhanh để tham dự đến quá khứ gần đây, trong: NIPS, trang 4331–4339..

[68] Ç. Gülçehre, S. Chandar, K. Cho, Y. Bengio, Máy xử lý thần kinh động với các sơ đồ địa chỉ cứng và mềm, CoRR abs/1607.00036 (2016).

[69] M. Daniluk, T. Rocktäschel, J. Welbl, S. Riedel, Khoảng thời gian chú ý ngắn đến đáng thất vọng trong mô hình ngôn ngữ thần kinh, trong: ICLR (Poster), OpenReview.net, 2017..

[70] RJ Williams, Các thuật toán theo độ dốc thống kê đơn giản để học tăng cường kết nối, Mach. Học hỏi. 8 (1992) 229–256.

[71] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Mạng ép và kích thích, IEEE Trans. Mẫu Hậu Môn. Mach. Trí tuệ. 42 (2020) 2011–2023.

[72] J. Ba, V. Mnih, K. Kavukcuoglu, Nhận dạng nhiều đối tượng với sự chú ý trực quan, trong: Y. Bengio, Y. LeCun (Eds.), Hội nghị quốc tế lần thứ 3 về Đại diện học tập, ICLR 2015, San Diego, CA , Hoa Kỳ, ngày 7–9 tháng 5 năm 2015, Kỷ yếu của Hội nghị..

[73] M. Jaderberg, K. Simonyan, A. Zisserman, và những người khác, Mạng biến áp không gian, trong: Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, trang 2017–2025..

[74] S. Chaudhari, G. Polatkan, R. Ramanath, V. Mithal, Một cuộc khảo sát chu đáo về các mô hình chú ý, CoRR abs/1904.02874 (2019)..

[75] J. Lu, J. Yang, D. Batra, D. Parikh, Câu hỏi-hình ảnh phân cấp cùng chú ý cho trả lời câu hỏi trực quan, trong: NIPS, trang 289–297..

[76] F. Fan, Y. Feng, D. Zhao, Mạng chú ý đa chi tiết để phân loại tình cảm ở cấp độ khía cạnh, trong: EMNLP, Hiệp hội Ngôn ngữ học tính toán, 2018, trang 3433–3442..

[77] W. Wang, SJ Pan, D. Dahlmeier, X. Xiao, Sự chú ý nhiều tầng kết hợp để cùng trích xuất các thuật ngữ khía cạnh và quan điểm, trong: AAAI, AAAI Press, 2017, trang 3316–3322.

[78] Y. Tay, AT Lưu, SC Hui, Hermitian mạng đồng chú ý để so khớp văn bản trong các miền bất đối xứng, trong: IJCAI, ijcai.org, 2018, trang 4425–4431..

[79] Q. Zhang, J. Fu, X. Liu, X. Huang, Mạng đồng chú ý thích ứng để nhận dạng thực thể được đặt tên trong các tweet, trong: AAAI, AAAI Press, 2018, trang 5674–5681.

[80] F. Nie, Y. Cao, J. Wang, C. Lin, R. Pan, Đề cập và đồng chú ý mô tả thực thể để định hướng thực thể, trong: AAAI, AAAI Press, 2018, trang 5908–5915.

[81] X. Li, K. Song, S. Feng, D. Wang, Y. Zhang, Mô hình mạng thần kinh đồng chú ý để phân tích nguyên nhân cảm xúc với nhận thức bối cảnh cảm xúc, trong: EMNLP, Hiệp hội Ngôn ngữ học tính toán, 2018, trang 4752–4757..

[82] Y. Tay, AT Lưu, SC Hui, J. Su, Ngư dân đọc từ vựng có cổng chú ý với sự đồng chú ý theo ngữ cảnh tương phản để phân loại tình cảm, trong: EMNLP, Hiệp hội Ngôn ngữ học tính toán, 2018, trang 3443–3453..

[83] B. Wang, K. Liu, J. Zhao, Mạng lưới thần kinh tái phát dựa trên sự chú ý bên trong để lựa chọn câu trả lời, trong: ACL (1), Hiệp hội Ngôn ngữ học Máy tính, 2016..

[84] L. Wu, F. Tian, L. Zhao, J. Lai, T. Liu, Chú ý từ để hiểu văn bản theo trình tự, trong: AAAI, AAAI Press, 2018, trang 5578–5585.

[85] J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, Chú ý sâu hơn đến việc kiểm duyệt nội dung người dùng lạm dụng, trong: EMNLP, Hiệp hội Ngôn ngữ học tính toán, 2017, trang 1125–1135..

[86] Z. Li, Y. Wei, Y. Zhang, Q. Yang, Mạng chuyển sự chú ý theo cấp bậc để phân loại tình cảm giữa các miền, trong: AAAI, AAAI Press, 2018, trang 5852–5859.

[87] X. Wang, RB Girshick, A. Gupta, K. He, Mạng thần kinh phi cục bộ, trong: CVPR, Hiệp hội máy tính IEEE, 2018, trang 7794–7803..

[88] C. Wu, F. Wu, J. Liu, Y. Huang, Biểu diễn mục và ngư ời dùng phân cấp với sự chú ý ba tầng để đư a ra khuyến nghị, trong: NAACL-HLT (1), Hiệp hội Ngôn ngữ học tính toán, 2019, tr. 1818–1826..

[89] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, Ứng dụng của mô hình chú ý hai cấp độ trong mạng lư ới thần kinh tích chập sâu để phân loại hình ảnh hạt mịn, trong: CVPR , Hiệp hội Máy tính IEEE, 2015, trang 842–850.

[90] J. Li, Z. Tu, B. Yang, MR Lyu, T. Zhang, Sự chú ý nhiều đầu với sự chính quy hóa bất đồng, trong: EMNLP, Hiệp hội Ngôn ngữ học tính toán, 2018, trang 2897–2903..

[91] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, C. Zhang, Disan: mạng lư ới tự chú ý định hư ớng để hiểu ngôn ngữ không có rnn/cnn, trong: AAAI, AAAI Press, 2018, trang 5446–5455..

[92] J. Du, J. Han, A. Way, D. Wan, Tự chú ý đến cấu trúc đa cấp để trích xuất quan hệ đư ợc giám sát từ xa, trong: EMNLP, Hiệp hội Ngôn ngữ học tính toán, 2018, trang 2216–2225..

[93] S. Venugopalan, M. Rohrbach, J. Donahue, RJ Mooney, T. Darrell, K. Saenko, Trình tự thành chuỗi – video thành văn bản, trong: Hội nghị quốc tế IEEE 2015 về Thị giác máy tính, ICCV 2015, Santiago, Chile , Ngày 7–13 tháng 12 năm 2015, Hiệp hội Máy tính IEEE, 2015, trang 4534–4542..

[94] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han, Q. Liu, Tạo chú thích hình ảnh thần kinh với đào tạo và tham khảo có trọng số, Cogn. Máy tính. 11 (2019) 763–777.

[95] X. Zhang, Q. Yang, Chuyển mạng chú ý phân cấp cho hệ thống hộp thoại tổng quát, Int. J. Tự động. Máy tính. 16 (2019) 720–736.

[96] R. Prabhavalkar, K. Rao, TN Sainath, B. Li, L. Johnson, N. Jaitly, So sánh các mô hình theo trình tự để nhận dạng giọng nói, trong: F. Lacerda (Ed.), Interspeech 2017, Hội nghị thư ờng niên lần thứ 18 của Hiệp hội Truyền thông Lời nói Quốc tế, Stockholm, Thụy Điển, ngày 20–24 tháng 8 năm 2017, ISCA, 2017, trang 939–943.

[97] S. Wang, J. Zhang, C. Zong, Học cách trình bày câu với sự hư ớng dẫn của sự chú ý của con ngư ời, trong: IJCAI, ijcai.org, 2017, trang 4137–4143..

[98] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, Mạng bộ nhớ đầu cuối, trong: NIPS, trang 2440–2448..

[99] J. Weston, S. Chopra, A. Bordes, Mạng bộ nhớ, trong: ICLR..

[100] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, R. Socher, Hỏi tôi bất cứ điều gì: Mạng bộ nhớ động để xử lý ngôn ngữ tự nhiên, trong: ICML, Tập 48 của Kỷ yếu Hội nghị và Hội thảo JMLR, JMLR.org, 2016, trang 1378–1387..

[101] M. Henaff, J. Weston, A. Szlam, A. Bordes, Y. LeCun, Theo dõi trạng thái thế giới với các mạng thực thể định kỳ, trong: Hội nghị quốc tế lần thứ 5 về Đại diện học tập, ICLR 2017, Toulon, Pháp, ngày 24 tháng 4 –26, 2017, Kỷ yếu theo dõi hội nghị, OpenReview.net, 2017..

[102] J. Gehring, M. Auli, D. Grangier, D. Yarats, YN Dauphin, Trình tự chuyển đổi sang học theo trình tự, trong: ICML, Tập 70 của Kỷ yếu Nghiên cứu Học máy, PMLR, 2017, trang 1243–1252 ..

[103] X. Li, W. Wang, X. Hu, J. Yang, Mạng hạt nhân chọn lọc, trong: Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, CVPR 2019, Long Beach, CA, Hoa Kỳ, ngày 16–20 tháng 6 năm 2019 , Tổ chức Thị giác Máy tính/ IEEE, 2019, trang 510– 519..

[104] MF Stollenga, J. Masci, FJ Gomez, J. Schmidhuber, Mạng sâu với sự chú ý có chọn lọc nội bộ thông qua kết nối phản hồi, trong: Z. Ghahramani, M. Welling, C. Cortes, ND Lawrence, KQ Weinberger (Eds.) , Những tiến bộ trong hệ thống xử lý thông tin thần kinh 27: Hội nghị thư ờng niên về hệ thống xử lý thông tin thần kinh 2014, ngày 8–13 tháng 12 năm 2014, Montreal, Quebec, Canada, trang 3545–3553..

[105] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Mạng chú ý kép để phân đoạn cảnh, trong: Hội nghị IEEE về Tầm nhìn Máy tính và Nhận dạng Mẫu, CVPR 2019, Long Beach, CA, USA, ngày 16–20 tháng 6 năm 2019, Tổ chức Thị giác Máy tính/ IEEE, 2019, trang 3146–3154..

[106] Y. Yuan, J. Wang, Ocnet: mạng ngữ cảnh đối tư ợng để phân tích cảnh, CoRR abs/ 1809.00916 (2018)..

[107] H. Zhao, Y. Zhang, S. Liu, J. Shi, CC Loy, D. Lin, J. Jia, Psanet: mạng chú ý không gian theo điểm để phân tích cảnh, trong: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Thị giác máy tính - ECCV 2018–15 Hội nghị Châu Âu, Munich, Đức, ngày 8–14 tháng 9 năm 2018, Kỷ yếu, Phần IX, tập 11213 của Ghi chú Bài giảng về Khoa học Máy tính, Springer, 2018, trang . 270–286..

[108] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: các mạng phi cục bộ gặp các mạng kích thích siết chặt và hơ n thế nữa, trong: Hội thảo quốc tế IEEE/CVF 2019 về Hội thảo Thị giác Máy tính, Hội thảo ICCV 2019, Seoul, Hàn Quốc (Miền Nam), ngày 27–28 tháng 10 năm 2019, IEEE, 2019, trang 1971–1980..

[109] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Mạng chú ý còn lại để phân loại hình ảnh, trong: Hội nghị IEEE 2017 về Thị giác máy tính và Nhận dạng mẫu, CVPR 2017, Honolulu, HI, Hoa Kỳ, ngày 21–26 tháng 7 năm 2017, Hiệp hội Máy tính IEEE, 2017, trang 6450–6458..

[110] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, F. Xu, Mạng phi cục bộ tổng quát nhỏ gọn, trong: S. Bengio, HM Wallach, H. Larochelle, K. Grauman , N. Cesa-Bianchi, R. Garnett (Eds.), Những tiến bộ trong hệ thống xử lý thông tin thần kinh 31: Hội nghị thư ờng niên về hệ thống xử lý thông tin thần kinh 2018, NeurIPS 2018, 3-8 tháng 12 năm 2018, Montréal, Canada, trang 6511-6520 ..

[111] S. Woo, J. Park, J. Lee, IS Kweon, CBAM: mô-đun chú ý khối tích chập, trong: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018–15 Hội nghị Châu Âu, Munich, Đức, ngày 8–14 tháng 9 năm 2018, Kỷ yếu, Phần VII, Tập 11211 của Ghi chú Bài giảng về Khoa học Máy tính, Springer, 2018, trang 3–19..

[112] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: sự chú ý xuyên suốt về phân đoạn ngữ nghĩa, trong: Hội nghị quốc tế IEEE/CVF 2019 về Thị giác máy tính, ICCV 2019, Seoul, Hàn Quốc (Miền Nam), ngày 27 tháng 10–ngày 2 tháng 11 năm 2019, IEEE, 2019, trang 603–612..

[113] H. Mi, Z. Wang, A. Ittycheriah, Sự chú ý đư ợc giám sát đối với dịch máy thần kinh, trong: J. Su, X. Carreras, K. Duh (Eds.), Kỷ yếu của Hội nghị năm 2016 về các phư ơ ng pháp thực nghiệm trong tự nhiên Xử lý ngôn ngữ, EMNLP 2016, Austin, Texas, Hoa Kỳ, ngày 1-4 tháng 11 năm 2016, Hiệp hội Ngôn ngữ học tính toán, 2016, trang 2283–2288..

[114] L. Liu, M. Utiyama, AM Finch, E. Sumita, Dịch máy thần kinh với sự chú ý có giám sát, trong: N. Calzolari, Y. Matsumoto, R. Prasad (Eds.), COLING 2016, Hội nghị quốc tế lần thứ 26 về Ngôn ngữ học tính toán, Kỷ yếu của Hội nghị: Tài liệu kỹ thuật, ngày 11–16 tháng 12 năm 2016, Osaka, Nhật Bản, ACL, 2016, trang 3093–3102..

[115] B. Yang, Z. Tu, DF Wong, F. Meng, LS Chao, T. Zhang, Lập mô hình địa phư ơ ng cho mạng lư ới tự chú ý, trong: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Kỷ yếu của Hội nghị về các phư ơ ng pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên năm 2018, Brussels, Bỉ, ngày 31 tháng 10 đến ngày 4 tháng 11 năm 2018, Hiệp hội Ngôn ngữ học tính toán, 2018, trang 4449–4458.

[116] SI Wang, CD Manning, Baselines và bigram: đơ n giản, tình cảm tốt và phân loại chủ đề, trong: Hội nghị thư ờng niên lần thứ 50 của Hiệp hội Ngôn ngữ học tính toán, Kỷ yếu của Hội nghị, ngày 8–14 tháng 7 năm 2012, Đảo Jeju, Hàn Quốc – Tập 2: Bài viết ngắn, Hiệp hội Ngôn ngữ học Máy tính, 2012, trang 90–94..

[117] AL Maas, RE Daly, PT Pham, D. Huang, AY Ng, C. Potts, Học vectơ từ để phân tích tình cảm, trong: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), The 49th Hội nghị thư ờng niên của Hiệp hội Ngôn ngữ học tính toán: Công nghệ ngôn ngữ con ngư ời, Kỷ yếu của Hội nghị, 19–24 tháng 6 năm 2011, Portland, Oregon, Hoa Kỳ, Hiệp hội Ngôn ngữ học máy tính, 2011, trang 142–150..

[118] B. Pang, L. Lee, Khai thác ý kiến và phân tích tình cảm, Found. Thông tin xu hư ớng Trở lại. 2 (2007) 1–135.

[119] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz, Một cách tiếp cận Bayes để lọc e-mail rác, trong: Học cách phân loại văn bản: Các bài viết từ hội thảo năm 1998, tập. 62, Madison, Wisconsin, trang 98–105..

[120] Y. Song, J. Wang, T. Jiang, Z. Liu, Y. Rao, Mạng bộ mã hóa chú ý để phân loại tình cảm mục tiêu, CoRR abs/1902.09314 (2019).

[121] A. Ambartsoumian, F. Popowich, Tự chú ý: một khối xây dựng tốt hơ n để phân tích tình cảm các bộ phân loại mạng thần kinh, trong: A. Balahur, SM Mohammad, V. Hoste, R. Klinger (Eds.), Kỷ yếu Hội thảo lần thứ 9 về các phư ơ ng pháp tiếp cận tính toán đối với tính chủ quan, tình cảm và phân tích truyền thông xã hội, WASSA@EMNLP 2018, Brussels, Bỉ, ngày 31 tháng 10 năm 2018, Hiệp hội Ngôn ngữ học tính toán, 2018, trang 130–139..

[122] D. Tang, B. Qin, T. Liu, Phân loại tình cảm cấp độ khía cạnh với mạng bộ nhớ sâu, trong: J. Su, X. Carreras, K. Duh (Eds.), Kỷ yếu Hội nghị về các phư ơ ng pháp thực nghiệm năm 2016 trong Xử lý ngôn ngữ tự nhiên, EMNLP 2016, Austin, Texas, Hoa Kỳ, ngày 1–4 tháng 11 năm 2016, Hiệp hội Ngôn ngữ học tính toán, 2016, trang 214–224..

[123] P. Zhu, T. Qian, Phân loại tình cảm ở cấp độ khía cạnh nâng cao với bộ nhớ phụ, trong: EM Bender, L. Derczynski, P. Isabelle (Eds.), Kỷ yếu của Hội nghị quốc tế lần thứ 27 về Ngôn ngữ học tính toán, COLING 2018, Santa Fe, New Mexico, Hoa Kỳ, ngày 20–26 tháng 8 năm 2018, Hiệp hội Ngôn ngữ học tính toán, 2018, trang 1077–1087..

[124] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, G. Hu, Mạng lư ới thần kinh chú ý quá mức để đọc hiểu, trong: R. Barzilay, M. Kan (Eds.), Kỷ yếu Hội nghị thư ờng niên lần thứ 55 của Hiệp hội Ngôn ngữ học tính toán, ACL 2017, Vancouver, Canada, ngày 30 tháng 7–ngày 4 tháng 8, Tập 1: Bài báo dài, Hiệp hội Ngôn ngữ học tính toán, 2017, trang 593–602..

[125] T. Mikolov, K. Chen, G. Corrado, J. Dean, Ư ớc tính hiệu quả các cách biểu diễn từ trong không gian vectơ , trong: Y. Bengio, Y. LeCun (Eds.), Hội nghị quốc tế lần thứ nhất về biểu diễn học tập, ICLR 2013, Scottsdale, Arizona, Hoa Kỳ, ngày 2–4 tháng 5 năm 2013, Kỷ yếu Hội thảo..

[126] J. Pennington, R. Socher, CD Manning, Glove: vectơ toàn cầu để biểu diễn từ, trong: A. Moschitti, B. Pang, W. Daelemans (Eds.), Kỷ yếu của Hội nghị năm 2014 về các phư ơ ng pháp thực nghiệm trong tự nhiên Xử lý ngôn ngữ, EMNLP 2014, ngày 25–29 tháng 10 năm 2014, Doha, Qatar, Cuộc họp của SIGDAT, Nhóm lợi ích đặc biệt của ACL, ACL, 2014, trang 1532–1543..

[127] ME Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Biểu diễn từ ngữ theo ngữ cảnh sâu sắc, trong: MA Walker, H. Ji, A. Stent (Eds.), Kỷ yếu Hội nghị năm 2018 của Chi hội Bắc Mỹ thuộc Hiệp hội Ngôn ngữ học Tính toán: Công nghệ Ngôn ngữ Con ngư ời, NAACL-HLT 2018, New Orleans, Louisiana, Hoa Kỳ, ngày 1–6 tháng 6 năm 2018, Tập 1 (Bài báo dài), Hiệp hội Ngôn ngữ học tính toán, 2018, trang 2227–2237..

[128] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: đào tạo trư ớc các máy biến áp hai chiều sâu để hiểu ngôn ngữ, trong: J. Burstein, C. Doran, T. Solorio (Eds.), Kỷ yếu Hội nghị năm 2019 của Chi hội Bắc Mỹ của Hiệp hội Ngôn ngữ học tính toán: Công nghệ ngôn ngữ con ngư ời, NAACL-HLT 2019, Minneapolis, MN, Hoa Kỳ, ngày 2–7 tháng 6,

2019, Tập 1 (Bài báo dài và ngắn), Hiệp hội Ngôn ngữ học tính toán, 2019, trang 4171–4186..

[129] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Cải thiện khả năng hiểu ngôn ngữ bằng cách đào tạo trư ớc mang tính tổng quát, 2018..

[130] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Mô hình ngôn ngữ là những ngư ời học đa nhiệm không đư ợc giám sát, OpenAI Blog 1 (2019) 9.

[131] TB Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, DM Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Mô hình ngôn ngữ là những ngư ời học ít lần, CoRR abs/2005.14165 (2020)..

[132] W. Liu, Z. Wang, N. Zeng, Y. Yuan, FE Alsaadi, X. Liu, Một trình tối ư u hóa bầy hạt ngẫu nhiên mới, Int. J. Mach. Học hỏi. Cybern. 12 (2021) 529–540.

[133] N. Zeng, Z. Wang, W. Liu, H. Zhang, K. Hone, X. Liu, Thuật toán tối ư u hóa bầy hạt chuyển mạch dựa trên vùng lân cận động, IEEE Trans. Cybern. (2020).

[134] W. Liu, Z. Wang, Y. Yuan, N. Zeng, K. Hone, X. Liu, Một công cụ tối ư u hóa nhóm hạt có trọng số thích ứng dựa trên chức năng sigmoid mới, IEEE Trans. Cybern. 51 (2021) 1085–1093.

[135] IU Rahman, Z. Wang, W. Liu, B. Ye, M. Zakarya, X. Liu, Thuật toán tối ư u hóa bầy hạt nhảy Markovian trạng thái n, IEEE Trans. Syst., Man, Cybern.: Syst. (2020).

[136] X. Luo, Y. Yuan, S. Chen, N. Zeng, Z. Wang, Phân tích hệ số tiềm ẩn đư ợc kết hợp tối ư u hóa bầy hạt chuyển tiếp vị trí, IEEE Trans. Kiến thức. Dữ liệu kỹ thuật số (2020).

[137] N. Zeng, D. Song, H. Li, Y. You, Y. Liu, FE Alsaadi, Một cơ chế cạnh tranh tích hợp thuật toán tối ư u hóa cá voi đa mục tiêu với tiến hóa vi sai, Neurocomputing 432 (2021) 170–182.

[138] J. Li, W. Monroe, D. Jurafsky, Tìm hiểu mạng lư ới thần kinh thông qua việc xóa biểu diễn, CoRR abs/1612.08220 (2016)..

[139] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, I. Titov, Phân tích sự tự chú ý của nhiều đầu: những cái đầu chuyên dụng đảm nhiệm việc nâng vật nặng, phần còn lại có thể cắt tỉa, trong: A. Korhonen, DR Traum, L. Màrquez (Eds.), Kỷ yếu Hội nghị lần thứ 57 của Hiệp hội Ngôn ngữ học Tính toán, ACL 2019, Florence, Ý, 28 tháng 7–2 tháng 8 năm 2019, Tập 1: Bài viết dài, Hiệp hội Ngôn ngữ học Tính toán , 2019, trang 5797–5808..

[140] Z. Dai, Z. Yang, Y. Yang, JG Carbonell, QV Le, R. Salakhutdinov, Transformer-xl: Các mô hình ngôn ngữ chú ý vư ợt ra ngoài ngữ cảnh có độ dài cố định, trong: A. Korhonen, DR Traum, L. Màrquez (Eds.), Kỷ yếu Hội nghị lần thứ 57 của Hiệp hội Ngôn ngữ học Tính toán, ACL 2019, Florence, Ý, 28 tháng 7– 2 tháng 8 năm 2019, Tập 1: Bài viết dài, Hiệp hội Ngôn ngữ học Tính toán , 2019, trang 2978–2988..

[141] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, L. Kaiser, Universal Transformers, trong: Hội nghị quốc tế lần thứ 7 về đại diện học tập, ICLR 2019, New Orleans, LA, Hoa Kỳ, ngày 6–9 tháng 5, 2019, OpenReview.net, 2019..

[142] Q. Guo, X. Qiu, P. Liu, Y. Shao, X. Xue, Z. Zhang, Star-transformer, trong: J. Burstein, C. Doran, T. Solorio (Eds.), Kỷ yếu Hội nghị năm 2019 của Chi hội Bắc Mỹ của Hiệp hội Ngôn ngữ học tính toán: Công nghệ ngôn ngữ con ngư ời, NAACL-HLT 2019, Minneapolis, MN, Hoa Kỳ, ngày 2–7 tháng 6 , 2019, Tập 1 (Bài báo dài và ngắn), Hiệp hội Ngôn ngữ học tính toán, 2019, trang 1315–1325..

[143] X. Zhu, D. Cheng, Z. Zhang, S. Lin, J. Dai, Một nghiên cứu thực nghiệm về cơ chế chú ý không gian trong các mạng sâu, trong: Hội nghị quốc tế IEEE/CVF 2019 về Thị giác máy tính, ICCV 2019, Seoul , Korea (South), ngày 27 tháng 10–ngày 2 tháng 11 năm 2019, IEEE, 2019, trang 6687–6696..

[144] Y. Tay, D. Bahri, D. Metzler, D. Juan, Z. Zhao, C. Zheng, Synthesizer: xem xét lại việc tự chú ý trong các mô hình máy biến áp, CoRR abs/2005.00743 (2020)..

[145] YH Tsai, S. Bai, M. Yamada, L. Morency, R. Salakhutdinov, Mỗ xẻ máy biến áp: Một sự hiểu biết thống nhất về sự chú ý của máy biến áp qua lăng kính hạt nhân, trong: K. Inui, J. Jiang, V. Ng , X. Wan (Eds.), Kỷ yếu của Hội nghị năm 2019 về các phư ơng pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên và Hội nghị chung quốc tế lần thứ 9 về xử lý ngôn ngữ tự nhiên, EMNLP-IJCNLP 2019, Hồng Kông, Trung Quốc, ngày 3–7 tháng 11 năm 2019, Hiệp hội cho Ngôn ngữ học tính toán, 2019, trang 4343–4352..

[146] A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, Máy biến áp là rnns: Máy biến áp tự hồi quy nhanh với chú ý tuyến tính, CoRR abs/2006.16236 (2020).

[147] C. Sen, T. Hartvigsen, B. Yin, X. Kong, EA Rundensteiner, Bản đồ chú ý của con ngư ời để phân loại văn bản: Con ngư ời và mạng lư ới thần kinh có tập trung vào cùng một từ không?, trong: D. Jurafsky, J. Chai , N. Schluter, JR Tetreault (Biên tập), Kỷ yếu Hội nghị thư ờng niên lần thứ 58 của Hiệp hội Ngôn ngữ học tính toán, ACL 2020, Trực tuyến, ngày 5–10 tháng 7 năm 2020, Hiệp hội Ngôn ngữ học tính toán, 2020, trang 4596–4608..

Zhaoyang Niu nhận bằng Cử nhân Khoa học Dữ liệu và Kỹ thuật Phần mềm của Đại học Thanh Đảo, Thanh Đảo, Trung Quốc vào năm 2019. Hiện anh đang học lấy bằng Cử nhân. bằng Công nghệ Máy tính tại Đại học Hải dư ơng Trung Quốc, Thanh Đảo, Trung Quốc. Mối quan tâm nghiên cứu của ông bao gồm thị giác máy tính, học sâu và cơ chế chú ý.

Guoqiang Zhong nhận bằng Cử nhân Toán học tại Đại học Sư phạm Hà Bắc, Thạch Gia Trang, Trung Quốc, bằng Thạc sĩ về Nghiên cứu Hoạt động và Điều khiển học tại Đại học Công nghệ Bắc Kinh (BJUT), Bắc Kinh, Trung Quốc và bằng Tiến sĩ. bằng về Nhận dạng Mẫu và Hệ thống Thông minh của Viện Tự động hóa, Viện Hàn lâm Khoa học Trung Quốc (CASIA), Bắc Kinh, Trung Quốc, lần lư ợt vào các năm 2004, 2007 và 2011.

Từ tháng 10 năm 2011 đến tháng 7 năm 2013, ông là Nghiên cứu sinh sau tiến sĩ của Phòng thí nghiệm Synchromedia về Truyền thông Đa phư ơng tiện trong Telepresence, Đại học Quebec, Montreal, Canada. Từ tháng 3 năm 2014 đến tháng 12 năm 2020, ông là phó giáo sư tại Khoa Khoa học và Công nghệ Máy tính, Đại học Hải dư ơng Trung Quốc, Thanh Đảo, Trung Quốc. Kể từ tháng 1 năm 2021, ông là giáo sư chính thức tại Khoa Khoa học và Công nghệ Máy tính, Đại học Hải dư ơng Trung Quốc. Ông đã xuất bản 4 cuốn sách, 4 chư ơng sách và hơn 80 tài liệu kỹ thuật trong lĩnh vực trí tuệ nhân tạo, nhận dạng mẫu, học máy và thị giác máy tính. Mối quan tâm nghiên cứu của ông bao gồm nhận dạng mẫu, học máy và thị giác máy tính. Ông đã từng là Chủ tịch/thành viên PC/ngư ời đánh giá cho nhiều hội nghị quốc tế và các tạp chí hàng đầu, chẳng hạn như IEEE TNNLS, IEEE TKDE, IEEE TCSVT, Nhận dạng mẫu, Hệ thống dựa trên tri thức, Máy tính thần kinh, ACM TKDD, AAAI, AISTATS, ICPR, IJCNN, ICONIP và ICDAR. Ông đã đư ợc một số tạp chí trao giải là nhà phê bình xuất sắc, chẳng hạn như Nhận dạng mẫu, Hệ thống dựa trên tri thức, Điện toán thần kinh và Nghiên cứu hệ thống nhận thức. Anh đã giành đư ợc Giải thư ởng Bài báo hay nhất của BICS2019 và Giải thư ởng Nhà nghiên cứu trẻ APNNS. Ông là thành viên của ACM, IEEE, IAPR, APNNS và CCF, thành viên ủy ban chuyên môn của CAAI-PR, CAA-PRMI và CSIG-DIAR, đồng thời là ủy viên của Hiệp hội Trí tuệ nhân tạo Sơ n Đông.

Hui Yu là Giáo sư tại Đại học Portsmouth, Vư ơng quốc Anh. Giáo sư Yu nhận bằng Tiến sĩ tại Đại học Brunel Lon-don. Ông từng làm việc tại Đại học Glasgow trư ớc khi chuyển đến Đại học Portsmouth. Mối quan tâm nghiên cứu của ông bao gồm các phư ơng pháp và phát triển thực tế về thị giác, học máy và AI với các ứng dụng tư ơng tác giữa ngư ời và máy, thực tế ảo và tăng cư ờng, robot và nét mặt 4D.

Ông là Phó biên tập của tạp chí IEEE Transactions on Human-Machine Systems và Neurocomputing.