Hướng tới sự thống nhất có nền tảng và

NGỮ NGHĨA PHÂN PHỐI SỬ DỤNG MÔ HÌNH TỪ-NHƯ-PHÂN LOẠI CỦA TỪ VỰNG NGỮ NGHĨA

qua

Stacy Đen

Một luận án

nộp một phần để hoàn thành

của các yêu cầu về mức độ

Thạc sĩ Khoa học Máy tính

Đại học Boise State

Tháng 8 năm 2020

khoảng năm 2020

Stacy Đen

TRƯỜNG CAO ĐẲNG ĐẠI HỌC BANG BOISE

ỦY BAN QUỐC PHÒNG VÀ PHÊ DUYỆT ĐỌC CUỐI CÙNG

của luận án được nộp bởi

Stacy Đen

Tiêu đề luận án: Hướng tới thống nhất ngữ nghĩa cơ bản và phân phối bằng cách sử dụng Mô hình từ ngữ như phân loại của ngữ nghĩa từ vựng

Ngày thi vấn đáp cuối kỳ : Ngày 24 tháng 4 năm 2020

Những cá nhân sau đây đã đọc và thảo luận luận án do sinh viên Stacy nộp Đen, và họ đã đánh giá bài thuyết trình và phản hồi các câu hỏi trong phần cuối kỳ thi vấn đáp. Họ thấy rằng sinh viên đã vượt qua kỳ thi vấn đáp cuối cùng.

Casey Kennington, Tiến sĩ	Chủ tịch, Ủy ban giám sát
Francesca Spezzano, Tiến sĩ	Thành viên, Ủy ban giám sát
Tiến sĩ Sole Pera	Thành viên, Ủy ban giám sát

Sự chấp thuận đọc cuối cùng của luận án đã được chấp thuận bởi Casey Kennington, Tiến sĩ, Chủ tịch Ủy ban giám sát. Luận án đã được Hội đồng sau đại học chấp thuận

Trường cao đẳng.

dành tặng cho Greg, James, Heather và Michael

LỜI CẢM ƠN

Tôi muốn bày tỏ lòng biết ơn chân thành của tôi đối với cố vấn của tôi, Tiến sĩ Casey Kennington, vì sự hỗ trợ và hướng dẫn của anh ấy trong suốt quá trình hoàn thành luận án này. Tôi cũng muốn cảm ơn Daniele Moro vì sự hiểu biết sâu sắc và hỗ trợ có giá trị của anh ấy trong một số khía cạnh của tác phẩm này.

Cuối cùng, tôi muốn cảm ơn gia đình tôi, và đặc biệt là chị gái tôi, vì tất cả tình yêu thương của họ và sự động viên trong suốt quá trình học tập của tôi.

TÓM TẮT

Các hệ thống tự động sử dụng ngôn ngữ, chẳng hạn như trợ lý cá nhân, cần một số phương tiện biểu diễn từ ngữ sao cho 1) biểu diễn có thể tính toán được và 2) nắm bắt hình thức và ý nghĩa. Những tiến bộ gần đây trong lĩnh vực ngôn ngữ tự nhiên quá trình xử lý đã dẫn đến những cách tiếp cận hữu ích để biểu diễn giá trị trung bình của từ có thể tính toán được ings. Trong luận án này, tôi xem xét hai cách tiếp cận như vậy: nhúng phân phối và mô hình cơ bản. Nhúng phân phối được biểu diễn dưới dạng chiều cao vectơ; các từ có ý nghĩa tương tự có xu hướng tập trung lại với nhau trong không gian nhúng. Nhúng có thể dễ dàng học được bằng cách sử dụng một lượng lớn dữ liệu văn bản. Tuy nhiên, nhúng thiếu kiến thức về "thế giới thực"; ví dụ, kiến thức về việc xác định màu sắc hoặc các đối tượng khi chúng xuất hiện. Ngược lại với nhúng, các mô hình cơ bản học một bản đồ giữa ngôn ngữ và thế giới vật chất, chẳng hạn như thông tin trực quan trong hình ảnh. Tuy nhiên, các mô hình có cơ sở có xu hướng chỉ tập trung vào việc lập bản đồ giữa ngôn ngữ và thế giới vật chất và thiếu kiến thức có thể thu được từ xem xét thông tin trừu tượng có trong văn bản.

Trong luận án này, tôi đánh giá wac2vec, một mô hình kết hợp giữa cơ sở và ngữ nghĩa phân phối để hướng tới việc tận dụng sức mạnh tương đối của cả hai, và sử dụng phân tích thực nghiệm để khám phá xem wac2vec có bổ sung thông tin ngữ nghĩa hay không đến các nhúng truyền thống. Bắt đầu với mô hình từ-như-phân-loại (WAC) của ngữ nghĩa có cơ sở, tôi sử dụng một kho lưu trữ hình ảnh lớn và các từ khóa đã được sử dụng để lấy lại những hình ảnh đó. Từ mô hình cơ sở, tôi trích xuất hệ số phân loại

các cients như nhúng vectơ cấp độ từ (do đó, wac2vec), sau đó kết hợp chúng với em-
các lớp phủ từ các biểu diễn từ phân phối. Tôi chỉ ra rằng kết hợp các lớp phủ
nhúng với nhúng truyền thống dẫn đến hiệu suất được cải thiện trong hình ảnh
nhiệm vụ, chứng minh tính khả thi của việc sử dụng mô hình wac2vec để làm giàu truyền thống
nhúng và cho thấy wac2vec cung cấp thông tin ngữ nghĩa quan trọng
những nhúng này không có tính riêng biệt.

MỤC LỤC

DANH SÁCH CÁC BẢNG

DANH SÁCH CÁC HÌNH ẢNH

DANH SÁCH CÁC TỪ VIẾT TẮT

NLP – Xử lý ngôn ngữ tự nhiên

AoA – Thời đại của sự tiếp thu

WAC – Từ ngữ như một công cụ phân loại

CNN – Mạng nơ-ron tích chập

CHƯƠNG 1

GIỚI THIỆU

1.1 Ý nghĩa được xác định và phân phối

Việc thể hiện một số loại xấp xỉ ngữ nghĩa của ngôn ngữ là điều cần thiết trong bất kỳ nhiệm vụ tự động sử dụng ngôn ngữ tự nhiên, bao gồm dịch máy, giọng nói phiên âm, và tìm kiếm trên web, trong số những thứ khác. Mặc dù nó tương đối tự động để con người có thể tiếp thu được ý nghĩa của từ ngữ, tiếp thu và biểu diễn ngữ nghĩa ý nghĩa trong máy tính vẫn là một thách thức chưa được giải quyết. Ứng dụng hiện tại tiếp cận ngữ nghĩa không có phương pháp học tập và biểu diễn toàn diện ngữ nghĩa có tính đến tất cả các khía cạnh của ý nghĩa ngữ nghĩa của một từ. Có một số cách tiếp cận cạnh tranh nhưng có khả năng bổ sung cho ngữ nghĩa: chính thức ngữ nghĩa, ngữ nghĩa phân phối và ngữ nghĩa cơ bản. Luận văn này tập trung vào đang tiến tới sự thống nhất của hai điều sau: phân phối, đã chứng kiến thành công trong những năm gần đây, và có cơ sở, mà chúng tôi cho rằng cung cấp thông tin quan trọng ngữ nghĩa phân phối thiếu. Bao gồm ngữ nghĩa hình thức nằm ngoài phạm vi của công việc.

Trong ngữ nghĩa phân phối, ý nghĩa của một thuật ngữ ngôn ngữ dựa trên cách rằng nó được phân phối giữa các thuật ngữ khác. Nói cách khác, một từ xuất hiện trong một ngữ cảnh tương tự như một từ khác có thể có ý nghĩa ngữ nghĩa tương tự. Một com- cách tiếp cận duy nhất đối với ngữ nghĩa phân phối là biểu diễn một từ như một từ có chiều cao

vector, được gọi là nhúng, chẳng hạn như Word2vec [32]. Ngược lại, ngữ nghĩa có cơ sở
cố gắng mô hình hóa cách thức ý nghĩa của một từ dựa trên nhận thức về thế giới
(ví dụ như hình ảnh hoặc âm thanh). Ví dụ, mọi người biết màu đỏ có nghĩa là gì vì họ có
nhìn thấy những vật thể được gọi là màu đỏ; họ không học được ý nghĩa ngữ nghĩa của nó bằng cách đọc
về nó trong từ điển.

Cả hai lý thuyết đều có những ưu điểm riêng: ví dụ, nhúng phân phối có thể
có thể dễ dàng được đào tạo trên văn bản và các mô hình cơ bản hữu ích trong các nhiệm vụ phụ thuộc
về nhận thức, chẳng hạn như robot và xe tự lái; những lợi thế này được thảo luận
sâu hơn sau này trong luận án này. Cả hai cũng đưa ra những giả định khiến họ
không có khía cạnh quan trọng của ý nghĩa: các mô hình ngữ nghĩa có cơ sở thường cho rằng
sự độc lập của các từ với nhau, bỏ qua ngữ cảnh từ vựng và phân phối
ngữ nghĩa cho rằng tất cả các từ đều có tính phân phối và bỏ qua thực tế là các nghĩa
của nhiều từ có một số loại neo trong thế giới thực. Bất kỳ ngôn ngữ tự nhiên nào
nhiệm vụ xử lý có những từ được gắn chặt về mặt vật lý có khả năng không thể
rút ra thông tin quan trọng (và kết quả là có khả năng hoạt động kém hơn)
khi tiếp cận từ góc độ phân phối thuần túy.

Để đạt được một mô hình thống nhất, chúng tôi tận dụng mô hình từ-làm-phân-loại (WAC)
[17]. WAC là một mô hình ngữ nghĩa có cơ sở đã được sử dụng trong nhiều
nhiệm vụ, chẳng hạn như giải quyết tham chiếu trong cuộc đối thoại với rô-bốt, nhưng tôi giải thích bên dưới cách
WAC có thể tạo ra các nhúng cấp độ từ mà sau đó chúng ta kết hợp với các nhúng hiện có
được đào tạo trên văn bản.

Mục tiêu của luận án này là tiến hành một cuộc khám phá thực nghiệm về các cấp độ từ ngữ này
Nhúng WAC. Cụ thể, mục tiêu là khám phá xem có nên sử dụng WAC làm
một nhúng có cơ sở làm phong phú thêm các nhúng chỉ có văn bản. Để thực hiện điều này
Mục tiêu, chúng tôi kiểm tra các nhúng WAC, mà chúng tôi gọi là wac2vec, trên một loạt các tác vụ NLP:

đối thoại trực quan, được chọn vì sử dụng ngôn ngữ có căn cứ (một nhiệm vụ mà việc thêm WAC nên cải thiện hiệu suất), cùng với việc phân cụm cụm từ, nhận dạng thực thể được đặt tên, và sự tương đồng của từ, các nhiệm vụ NLP phổ biến cung cấp cái nhìn sâu sắc hơn về những gì mô hình học, chẳng hạn như hiểu cú pháp (một cái gì đó nhúng chỉ có văn bản nên hoạt động tốt tại), và nhúng có xu hướng nhóm lại về mặt ngữ nghĩa hay không. đánh giá cho thấy WAC đóng góp thông tin ngữ nghĩa, nhưng không phải cú pháp, chỉ ra rằng nó sẽ hữu ích trong các nhiệm vụ cơ bản và nó sẽ cần phải được kết hợp với nhúng phân phối, không được sử dụng riêng lẻ, để sử dụng hiệu quả nhất. Hơn nữa, mô hình này có thể được sử dụng như một bộ phân loại cơ bản và một bộ nhúng, một bước quan trọng hướng tới một mô hình thống nhất có thể đại diện cho cả cơ sở và ý nghĩa phân phối.

Đối với phần còn lại của luận án này, trước tiên tôi trình bày tuyên bố luận án của mình, và sau đó thảo luận công trình liên quan đã được thực hiện về ngữ nghĩa phân phối và ngữ nghĩa cơ bản, và sự thống nhất của hai. Trong các chương sau, tôi mô tả dữ liệu được sử dụng trong công việc, và sau đó là mô hình và đánh giá của chúng tôi để trả lời câu luận đề. Cuối cùng, Tôi kết thúc bằng phần kết luận và công việc trong tương lai.

1.2 Luận đề

Hai lý thuyết ngữ nghĩa nổi bật, phân phối và cơ sở, mô tả hai những cách mà ngữ nghĩa của ngôn ngữ có thể được học và thể hiện; mỗi lý thuyết có điểm mạnh và điểm yếu, và mỗi điểm đều có thể biểu diễn ý nghĩa ngữ nghĩa mà khác thì không. Câu hỏi của chúng tôi trong luận án này là, “Có bao gồm nhúng mà đã được học bằng cách sử dụng thông tin trực quan làm phong phú thêm các nhúng chỉ có văn bản?” Chúng tôi đưa ra giả thuyết rằng, tận dụng mô hình WAC, các hệ số từ một bộ phân loại mạng nơ-ron được đào tạo

để hoạt động như một biểu diễn có cơ sở của ý nghĩa ngữ nghĩa ở cấp độ từ có thể là được sử dụng cùng với nhúng phân phối truyền thống, thêm ý nghĩa ngữ nghĩa không được nắm bắt bởi cách tiếp cận phân phối, và điều này sẽ dẫn đến cải thiện hiệu suất trên một số tác vụ xử lý ngôn ngữ tự nhiên (NLP) được chọn để thực nghiệm khám phá nơi WAC đóng góp thông tin ngữ nghĩa.

CHƯƠNG 2

BỐI CẢNH VÀ CÔNG VIỆC LIÊN QUAN

2.1 Bối cảnh

Trong công trình của mình, chúng tôi xem xét cách thức hai phương pháp tiếp cận ngữ nghĩa, có cơ sở và phân phối quốc tế, có thể được thống nhất. Trong phần này, chúng tôi trình bày công trình nghiên cứu cơ bản về hai các lý thuyết ngữ nghĩa, ngoài việc thảo luận về các từ cụ thể và trừu tượng, có liên quan đến hai lý thuyết này, cũng như đến công trình này. Tóm tắt các điểm mạnh và điểm yếu của các lý thuyết cơ bản và phân phối có thể được tìm thấy trong Bảng 2.1.

2.1.1 Từ ngữ cụ thể và trừu tượng

Điều quan trọng là trước tiên phải thảo luận về các từ cụ thể và trừu tượng. Mặc dù chúng tôi không đào tạo mô hình của chúng tôi về tính cụ thể hoặc trừu tượng của từ ngữ trực tiếp, chúng gắn liền với cách chúng tôi xây dựng tập dữ liệu của mình và liên kết với hai lý thuyết ngữ nghĩa - có cơ sở và phân phối, tương ứng.

Từ cụ thể, như được định nghĩa trong [21], là “từ dùng để chỉ các vật thể có thể hình dung được và hành động.” Nói cách khác, chúng đề cập đến những thứ vật lý có thể được phát hiện bằng các giác quan, và mọi người có thể dễ dàng hình thành hình ảnh tinh thần về chúng [16]. “Đỏ” và “mèo” là hai từ có tính cụ thể cao.

Ngược lại, các từ trừu tượng không đề cập đến những thứ vật lý mà là những thứ cảm xúc. và trạng thái tinh thần, ý tưởng, v.v. [16]. Các từ trừu tượng không dễ dàng được khái niệm hóa

Bảng 2.1: Điểm mạnh và điểm yếu của lý thuyết ngữ nghĩa phân phối và lý thuyết ngữ nghĩa cơ bản.

Lý thuyết	Điểm mạnh	Điểm yếu
Phân phối	Dễ đào tạo, hiệu quả trong nhiều nhiệm vụ, xem xét bối cảnh từ vựng	Không xem xét cách ngữ nghĩa của nhiều từ dựa trên nhận thức, khó diễn giải
Kết nối nối đất	của các khía cạnh ngôn ngữ với nhận thức và hiện thân, xem xét bối cảnh vật lý, đôi khi có thể diễn giải được	Thiếu thông tin phân loại và trừu tượng về ngôn ngữ đang sử dụng; không xem xét bối cảnh từ vựng trừu tượng

về mặt thị giác như những từ ngữ cụ thể; chúng không thể được phát hiện bằng các giác quan vật lý, nhưng thay vào đó phải được mô tả bằng những từ khác. Một số ví dụ về các từ trừu tượng là "dân chủ" và "phép tính".

Bây giờ chúng ta thảo luận về ngữ nghĩa phân phối (nắm bắt tốt nhất ngữ nghĩa của (từ trừu tượng), lý thuyết ngữ nghĩa đầu tiên trong hai lý thuyết được sử dụng trong luận án này.

2.1.2 Ngữ nghĩa phân phối

"Bạn sẽ biết một từ bằng cách kết bạn với nó" [13] là một khái niệm hình thành cơ sở đối với giả thuyết phân phối của Firth, cung cấp nền tảng cho cách thức các phương pháp tiếp cận phân phối, bao gồm nhúng (được thảo luận sau), được mô hình hóa. Sớm làm việc trên các mô hình ngữ nghĩa phân phối từ và công ty của họ bằng cách sử dụng từ đồng ma trận xuất hiện (tức là số lần từ X xuất hiện trong ngữ cảnh của từ Y dựa trên các cửa sổ được xác định trước về khoảng cách giữa hai từ trong văn bản) như được giải thích bởi Turney và Pantel [45].

Các công trình gần đây hơn đã cách mạng hóa cách ngôn ngữ có thể được thể hiện trên ma-Trung Quốc thông qua việc sử dụng nhúng. Nhúng là các vectơ có chiều cao xấp xỉ thông tin ngữ nghĩa của các từ và cụm từ. Các mô hình nổi tiếng

bao gồm Word2vec [32], GloVe [34], fastText [6], ELMo [35], Flair [3] và BERT [11].
Nhúng có lợi thế là chúng dễ dàng được đào tạo trên văn bản có sẵn
dữ liệu và có thể được sử dụng trong một số ứng dụng, bao gồm mã hóa dữ liệu văn bản
đầu vào cho mạng nơ-ron và các mô hình học máy khác. Hơn nữa, chúng là
có khả năng thu thập thông tin phân loại.

Word2vec là một công nghệ nhúng ban đầu được Mikolov và cộng sự giới thiệu vào năm 2013 [32] và cho thấy các mô hình dựa trên dự đoán sử dụng các hệ số của lớp ẩn có thể được coi là biểu diễn vectơ (tức là nhúng) của một từ; một ý tưởng từ mà chúng tôi lấy cảm hứng từ cách tiếp cận của mình, được mô tả dưới đây. Trong Word2vec, mô hình được đào tạo bằng cách cố gắng dự đoán một từ dựa trên ngữ cảnh của nó hoặc ngược lại. fastText là một phần mở rộng của Word2vec trong đó mỗi từ được tạo thành từ các ký tự n-gram và vectơ cho một từ là vectơ trung bình của các phần n-gram của nó [6]. Điều này cho phép fastText tạo ra các nhúng cho các từ mà nó chưa từng thấy trước đây. Trong GloVe, thay vì là sản phẩm phụ của mô hình, nhúng được tối ưu hóa cụ thể bằng sử dụng phân tích ma trận trong mô hình hồi quy log-bilinear, không giống như Word2vec, chứa thông tin từ vựng toàn cầu [34].

Nhúng Flair được lấy từ các trạng thái bên trong của một mô hình ngôn ngữ đã được đào tạo ở cấp độ nhân vật. Flair hoạt động bằng cách truyền tải toàn bộ một câu vào mô hình ngôn ngữ, từ đó các trạng thái nội bộ được lấy lại [3]. Như vậy, Nhúng Flair hoạt động ở cấp độ câu và nhúng cho một từ sẽ thay đổi tùy thuộc vào ngữ cảnh của nó. Gần đây hơn và mạnh mẽ hơn nhiều là BERT, một lớp biến áp hai chiều đã nâng cao trình độ nghệ thuật cho nhiều NLP- nhiệm vụ liên quan [11]. Mô hình được đào tạo trước bằng cách sử dụng BooksCorpus [47] và tiếng Anh Wikipedia, và sau đó được tinh chỉnh thêm cho từng nhiệm vụ. Giống như Flair, BERT hoạt động ở cấp độ câu và có khả năng nắm bắt các bối cảnh khác nhau cho ý nghĩa của một từ - ví dụ

Ví dụ, trong trường hợp “Cô ấy bị mất điện thoại di động” và “Virus tấn công tế bào màng.”

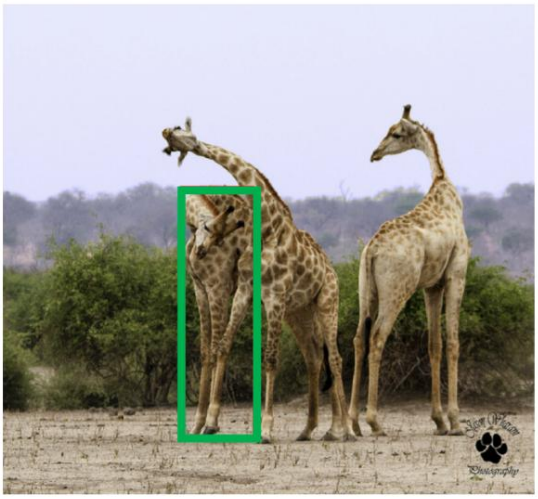
Mặc dù có những tiến bộ trong nhúng, vẫn còn một số thách thức. Như dis-được Herbelot [14] chỉ trích, các tập hợp văn bản có thể không bao gồm thông tin ngữ nghĩa quan trọng ví dụ, “Mèo có hai mắt” là một câu không thể viết được trong văn bản, vì bất kỳ ai biết mèo là gì đều thấy rõ ràng (một có thể lập luận rằng thông tin này là có cơ sở, chứ không phải là phân phối). thiếu thông tin văn bản này có nghĩa là các mô hình nhúng không thể học được khía cạnh ngữ nghĩa cụ thể này của từ cat. L¨ucking et al. [29] chỉ ra một số điểm yếu của ngữ nghĩa phân phối, bao gồm cả việc không có khả năng học một số loại biểu đạt ngôn ngữ (ví dụ: chỉ mục, tên riêng và từ wh). Một số thách thức khác là các vector nhúng khó diễn giải và các mô hình phân phối có xu hướng yêu cầu một lượng lớn dữ liệu đào tạo để học đại diện hiệu quả.

Điều quan trọng đối với công việc của tôi trong luận án này là giả định sai lầm rằng nhúng tạo ra: kiến thức ngữ nghĩa chỉ được rút ra từ ngữ cảnh từ vựng; nghĩa là, các từ khác trong văn bản [24, 14, 5]. Mặc dù nhúng thể hiện sự xấp xỉ tốt hơn về ngữ nghĩa thông tin cho các từ hơn các phương pháp trước đây chỉ sử dụng chuỗi từ đó hoặc tần suất tương đối trong một văn bản, bất kỳ nhúng nào cũng không thể được sử dụng trong các nhiệm vụ mà nhận thức là cần thiết, chẳng hạn như xác định màu sắc hoặc hình dạng. Một mô hình nhúng có thể biết, ví dụ Ví dụ, từ đỏ tập hợp lại–hoặc gần nhau trong không gian đa chiều–với các từ khác từ ngữ về màu sắc, nhưng nó không thể nói liệu có một quả táo được đưa cho máy ảnh hay không là màu đỏ.

Để “mở mắt” cho việc học ngữ nghĩa, bây giờ chúng ta chuyển sang thực tế ngữ nghĩa.

2.1.3 Ngữ nghĩa cơ bản

Ngữ nghĩa cơ bản kết nối ngôn ngữ với nhận thức và hiện thân (và tốt nhất nắm bắt ngữ nghĩa của các từ cụ thể), một cái gì đó mà nhúng có truyền thống bị bỏ qua một cách nghiêm trọng. Ví dụ, trong lĩnh vực ngôn ngữ trong ngành robot, Chai et al. [8] mô tả một cách tiếp cận để xây dựng ngôn ngữ cơ bản với hành động của người máy, và Thomason et al. ngôn ngữ cơ sở với nhận thức của rô-bốt [44]. Chen et al. áp dụng nền tảng trong một nhiệm vụ lý luận không gian và điều hướng được đề xuất [9], trong đó một tác nhân phải tuân theo hướng dẫn điều hướng trong môi trường điều hướng trực quan.



Hình 2.1: Một hình ảnh và các biểu thức tham chiếu đi kèm từ tập dữ liệu RefCOCO.

Một ví dụ về nơi các từ được gắn với bối cảnh vật lý là trong các tập dữ liệu như Ref-COCO [46], được mô tả trong Hình 2.1. Ở đây các từ của hai biểu thức tham chiếu ("con hươu cao cổ bên trái" và "con hươu cao cổ đầu tiên bên trái") rõ ràng phụ thuộc vào những gì về mặt vật lý được mô tả trong bức ảnh.

Các công trình khác của Kennington & Schlangen trình bày một từ như một bộ phân loại (WAC) mô hình dựa trên các từ ngữ với các khía cạnh trực quan của các đối tượng [17]. Trong cách tiếp cận này, cá nhân các từ ngữ trực quan được đào tạo như các bộ phân loại có thể dự đoán mức độ "phù hợp" giữa từ và các đối tượng được mô tả bởi từ đó. Nghĩa là, cho một đối tượng, bộ phân loại cho một từ nhất định có thể dự đoán liệu đối tượng có phải là một ví dụ về từ đó hay không từ. Mô hình WAC thậm chí có thể học được ý nghĩa của từ cơ bản chỉ với một vài ví dụ đào tạo.

Giống như trường hợp nhúng, có một số thách thức trong lĩnh vực này của ngữ nghĩa có cơ sở. Đối với một, hầu hết các nghiên cứu tập trung vào nhận thức trực quan, tuy nhiên Ý nghĩa ngữ nghĩa của nhiều từ không chỉ dựa trên hình ảnh; những từ như mềm hoặc chạy được nghiền thành nhận thức xúc giác và khứu giác (tức là mùi). Được nghiền thành các mô hình ngữ nghĩa cũng bị hạn chế về khả năng biểu diễn các khái niệm trừu tượng, chẳng hạn như như nền dân chủ, không có bất kỳ hình ảnh trực quan nào được mô tả rõ ràng. Và không giống như nhúng, các mô hình này thường không có thông tin phân loại.

Các mô hình ngữ nghĩa có cơ sở đưa ra một giả định trái ngược với giả định- mà nhúng tạo ra: trong khi nhúng chỉ xem xét ngữ cảnh từ vựng, có cơ sở các mô hình chỉ xem xét bối cảnh vật lý - các từ thường được đào tạo mà không có bất kỳ kết nối nào với những từ khác xung quanh chúng.

Do đó, mục tiêu của công trình trình bày trong luận án của tôi là phát huy thế mạnh của phân phối (tức là nhúng) và ngữ nghĩa cơ bản cùng nhau theo cách như vậy để để tận dụng thế mạnh của cả hai, bằng cách sử dụng ngữ cảnh từ vựng từ phân phối mô hình ngữ nghĩa và bối cảnh vật lý của mô hình ngữ nghĩa có cơ sở.

2.2 Công trình liên quan

Một số bài báo đã thừa nhận nhu cầu kết hợp cả cơ sở và phân phối
thông tin chuyên ngành để tạo ra sự biểu diễn toàn diện hơn về ý nghĩa ngữ nghĩa;
như Bruni et al. đã nêu [7], kết hợp ngôn ngữ với hình ảnh "chạm vào
các khía cạnh khác nhau của ý nghĩa." Sau đây là thảo luận về những cách tiếp cận khác nhau này
để giải quyết vấn đề này.

Elia Bruni, Gemma Boleda, Marco Baroni và Nam-Khánh Trần tiến hành
một nghiên cứu ban đầu để tìm hiểu cách sử dụng thông tin trực quan để tạo ra các mô hình tốt hơn
của ý nghĩa từ [7]. Trích dẫn Louwerse [26], họ nói rằng "[sự xuất hiện đồng thời của từ]
các mô hình...chỉ dựa vào thông tin bằng lời nói, trong khi kiến thức ngữ nghĩa của con người cũng dựa vào
về kinh nghiệm phi ngôn ngữ và sự thể hiện...chủ yếu là về thông tin thu thập được
thông qua nhận thức." Bruni và cộng sự đã xây dựng và thử nghiệm trực quan, văn bản và đa phương thức
(kết hợp các mô hình trực quan và văn bản) [7]. Các biểu diễn văn bản được tạo ra bằng cách sử dụng
mô hình đồng hiện từ với các kích thước cửa sổ ngữ cảnh khác nhau, trong khi hình ảnh
thông tin được thể hiện bằng các đặc điểm hình ảnh cục bộ đơn giản. Họ thấy rằng trong khi
các mô hình trực quan thực hiện kém hơn các mô hình văn bản trong các nhiệm vụ ngữ nghĩa chung, chúng
tốt hoặc tốt hơn khi mô hình hóa ý nghĩa của các từ bằng các tương quan trực quan,
bao gồm trong một nhiệm vụ mà mô hình phải phân biệt giữa không theo nghĩa đen (tức là màu xanh lá cây
gỗ, rượu vang trắng) và sử dụng theo nghĩa đen (tức là khăn trắng, lông vũ đen) của những từ như vậy. Trong
nhiệm vụ này, các mô hình đa phương thức thực hiện tốt nhất từ hoàn toàn trực quan, hoàn toàn văn bản,
và các mô hình đa phương thức.

Angeliki Lazaridou, Elia Bruni và Marco Baroni trình bày cách tiếp cận
ngữ nghĩa dựa trên vectơ đa phương thức cho việc học nhiệm vụ không có cú đánh nào, trong đó hệ thống
được trình bày với một đối tượng chưa từng thấy trước đó và phải ánh xạ nó với biểu diễn ngôn ngữ

việc diễn đạt một từ [23]. Các tác giả “không nhằm mục đích làm phong phú thêm các biểu diễn từ với thông tin trực quan, mặc dù đây có thể là tác dụng phụ của cách tiếp cận của chúng tôi, nhưng chúng tôi giải quyết vấn đề tự động ánh xạ các đối tượng, như được mô tả trong hình ảnh, để các vectơ ngữ cảnh biểu diễn các từ tương ứng.” Trong tác phẩm này, họ lấy các tính năng cấp thấp từ hình ảnh, tạo ra các vectơ dựa trên văn bản bằng cách sử dụng sự đồng hiện của từ, và sử dụng các biểu diễn này trong một mạng nơ-ron đơn giản được đào tạo trên đặc điểm hình ảnh các vector và đưa ra một vector dạng văn bản.

Douwe Kiela và L´eon Bottou hướng đến mục tiêu cải thiện các biểu diễn từ đa phương thức bằng cách sử dụng học chuyển giao và trích xuất biểu diễn hình ảnh từ một tích chập mạng nơ-ron nhân tạo (CNN) [18]. Các biểu diễn trực quan này sau đó được kết hợp với các biểu diễn ngôn ngữ skip-gram. Kiela và Bottou phát hiện ra rằng không chỉ hiệu suất với các vectơ đặc trưng CNN tốt hơn so với túi-của- truyền thống phương pháp tiếp cận từ ngữ trực quan sử dụng các đặc điểm hình ảnh cục bộ, tất cả các biểu diễn đa phương thức rằng họ đã thử nghiệm vượt trội hơn các biểu diễn chỉ mang tính ngôn ngữ; thứ hai này phát hiện này phản ánh những gì đã được chứng minh trong các nghiên cứu trước đây.

Ryan Kiros, Ruslan Salakhutdinov và Richard S. Zemel đã trình bày một cách tiếp cận để tạo chú thích hình ảnh (một nhiệm vụ trong đó chú thích mô tả được tạo ra cho hình ảnh) sử dụng đường ống mã hóa-giải mã, với không gian nhúng đa phương thức sử dụng hình ảnh và văn bản, và một mô hình ngôn ngữ để giải mã các biểu diễn từ không gian này [20]. Đối với các biểu diễn, họ đã tạo ra các nhúng từ với Word2vec tiếp cận [32] và lấy lại các vectơ hình ảnh bằng cách sử dụng CNN. Biểu diễn hình ảnh mô tả được tạo ra bằng LSTM.

Tanmay Gupta, Alexander Schwing và Derek Hoiem đã giới thiệu một phương pháp học cách nhúng từ vào các sự kiện trực quan, sử dụng chú thích văn bản tuổi (nếu hai từ mô tả cùng một hình ảnh hoặc vùng hình ảnh, chúng cùng xuất hiện) [1]. Chúng

trích xuất bốn loại sự đồng hiện trực quan giữa các từ đối tượng và thuộc tính
(thuộc tính đối tượng, thuộc tính-thuộc tính, ngữ cảnh và siêu ẩn danh đối tượng). Mô hình của họ
là một phần mở rộng đa nhiệm của GloVe mã hóa tất cả bốn loại cặp từ thành một
vector, để tránh một vector dài mà tỷ lệ tuyến tính với số lượng
các loại đồng hiện diện. Vectơ trực quan này được nối với GloVe tương ứng
vectơ.

Jamie Ryan Kiros, William Chan và Geoffrey E. Hinton đã tạo ra một phương thức đa phương thức
mô hình ngôn ngữ thần kinh cho mục đích tạo chú thích hình ảnh [19]. Một trong những
những điểm chính trong nghiên cứu của họ là việc sử dụng các nhúng “làm nền tảng cho ngôn ngữ bằng cách sử dụng
'ảnh chụp nhanh' được trả về bởi một công cụ tìm kiếm hình ảnh.” Họ tuyên bố rằng “mặc dù tự nhiên thực sự
hiểu ngôn ngữ có thể đòi hỏi nhận thức hoàn toàn hiện thân, công cụ tìm kiếm cho phép chúng ta
để có được một hình thức gần như là nền tảng từ những 'bức ảnh chụp nhanh' có độ bao phủ cao về thế giới vật chất của chúng ta
được cung cấp bởi sự tương tác của hàng triệu người dùng.” Đối với mỗi từ trong một vốn từ vựng, họ
đã lấy top-k từ tìm kiếm hình ảnh của Google, sau đó lấy các vectơ hình ảnh cho
mỗi từ bằng cách truyền chúng qua CNN. Họ sử dụng cơ chế gating đa phương thức
lựa chọn giữa các nhúng cơ bản này và nhúng GloVe, tùy thuộc vào
về mức độ trực quan của từ này.

Jiasen Lu, Dhruv Batra, Devi Parikh và Stefan Lee mở rộng mô hình BERT
để tạo ra một mô hình cho nền tảng thị giác học được các kết nối giữa thị giác
và văn bản (ViLBERT) [28]. Thay vì học thông tin trực quan và ngôn ngữ
riêng biệt, mô hình của họ xử lý cả dữ liệu đầu vào trực quan và văn bản cùng một lúc
trong kiến trúc luồng kép. Mỗi luồng tương tác với luồng kia thông qua sự hợp tác
lớp biến đổi sự chú ý. Họ tinh chỉnh và đánh giá ViLBERT trên một số
các nhiệm vụ về thị giác và ngôn ngữ, thiết lập trạng thái hiện đại cho tất cả các nhiệm vụ đó.

2.2.1 So sánh các phương pháp tiếp cận

Công trình trước đây trong việc kết hợp ngữ nghĩa cơ bản với ngữ nghĩa phân phối sử dụng các đặc điểm hình ảnh cơ bản được trích xuất bằng cách sử dụng SIFT [27]. SIFT là một công nghệ phát hiện đặc điểm hình ảnh thuật toán lấy các đặc điểm cục bộ từ một số điểm chính giới hạn trên hình ảnh, thường được sử dụng để nhận dạng đối tượng hoặc cảnh – nó không phải là sự biểu diễn của toàn bộ hình ảnh. Cả Bruni et al. [7] và Lazaridou et al. [23] đều sử dụng phương pháp này để tạo ra biểu diễn trực quan. Các tác phẩm sau này truyền hình ảnh qua CNN để trích xuất tính năng vectơ [18, 20, 19, 28]. Sử dụng CNN cung cấp một biểu diễn hoàn chỉnh hơn về một hình ảnh, vì nó được lấy từ toàn bộ hình ảnh, không chỉ là các điểm được chọn. Trong cách tiếp cận, chúng tôi sử dụng EfficientNet, một kiến trúc CNN hiện đại vừa được công bố bởi Tan và Le [42]. Trong Gupta et al. [1], không có dữ liệu hình ảnh thô nào được sử dụng; thay vào đó, nó là được tham chiếu gián tiếp bằng chú thích mô tả vùng hình ảnh.

Khi biểu diễn văn bản, các tác phẩm cũ hơn sử dụng các mô hình đồng hiện [7, 23]. Gần đây hơn tác phẩm sử dụng nhúng ngôn ngữ như Word2vec và GloVe [18, 20, 1, 19, 28]. Chúng tôi theo dõi các phương pháp tiếp cận trước đó và sử dụng nhúng GloVe để biểu diễn các từ.

Nhiều cách tiếp cận trong lĩnh vực này chỉ đơn giản là nối kết biểu diễn văn bản với biểu diễn trực quan [7, 18, 1], mà chúng tôi cũng thực hiện ở đây. Lazaridou et al. thực hiện một cách tiếp cận khác nhau – họ học cách lập bản đồ đa phương thức để tìm ra ý nghĩa của từ đã được xác định bằng cách đào tạo một mạng nơ-ron lấy các vectơ đặc trưng hình ảnh làm đầu vào và đầu ra một vector dựa trên văn bản [23]. Các tác phẩm khác tận dụng cả thông tin văn bản và hình ảnh không tạo ra các nhúng mục đích chung mà thay vào đó là nhúng cụ thể cho một nhiệm vụ nhất định [20, 19]. Trong ViLBERT, mô hình được cho là sẽ được áp dụng cho các nhiệm vụ thị giác học; nó không tạo ra các nhúng mục đích chung [28].

Cách tiếp cận của chúng tôi thường dựa trên công trình chúng tôi đã thực hiện trong Moro et al., trong đó chúng tôi chứng minh
đã xây dựng một kiến trúc đơn giản cho WAC, kiến trúc này gợi ý tiềm năng sử dụng
các hệ số của mô hình được đào tạo như nhúng từ cơ bản, mặc dù quá trình khám phá của chúng tôi
của tiềm năng này bị giới hạn ở một nhiệm vụ NLP duy nhất, cho thấy kết quả hỗn hợp [33].
Chúng tôi thảo luận về mô hình WAC chi tiết hơn trong Phần 3.2. Chúng tôi đào tạo một tập hợp WAC
bộ phân loại sử dụng tập dữ liệu hình ảnh được lấy từ công cụ tìm kiếm hình ảnh
sử dụng vốn từ vựng của các từ và sử dụng CNN để trích xuất các tính năng từ những từ này
hình ảnh. Công trình của chúng tôi trong luận án này dựa trên mô hình đơn giản trong Moro et al. và
phân tích tính hữu ích của nó trong một số nhiệm vụ NLP. Chúng tôi chỉ ra tính khả thi của việc sử dụng
WAC như một nhúng được đào tạo trên thông tin trực quan để làm giàu các nhúng hiện có
giống như GloVe. Hơn nữa, theo hiểu biết của chúng tôi, không có cách tiếp cận nào khác có thể được sử dụng làm cơ sở
phân loại và như một nhúng (mặc dù trong luận án này, chúng tôi chỉ thử nghiệm nhúng,
như đã mô tả trong Chương 4; công trình trước đây đã chỉ ra hiệu quả của các bộ phân loại WAC
trong các nhiệm vụ có cơ sở), một bước quan trọng hướng tới việc thống nhất các nhiệm vụ có cơ sở và phân phối
ngữ nghĩa để đi đến một mô hình duy nhất có thể học và biểu diễn cả cơ sở và
thông tin phân phối.

CHƯƠNG 3

PHƯƠNG PHÁP

Trong luận án này, chúng tôi theo dõi công trình ban đầu được thực hiện trong Moro et al. và sử dụng phương pháp WAC được đề xuất bởi Kennington và Schlangen để tạo ra các nhúng từ có cơ sở, hiệu quả mà sau này chúng tôi thử nghiệm trong một số thí nghiệm NLP [33, 17].

Trong các phần của chương này, trước tiên chúng ta thảo luận về dữ liệu được sử dụng trong cả quá trình đào tạo của mô hình của chúng tôi và trong các thí nghiệm sau đó. Sau đó, chúng tôi mô tả mô hình WAC và các nhúng được trích xuất từ bộ phân loại WAC, cách chúng tôi đào tạo Bộ phân loại WAC và điều chỉnh thông số tiếp theo.

3.1 Dữ liệu

3.1.1 Từ vựng và Bộ dữ liệu

Đối với mục đích của luận án này, chúng tôi xây dựng một tập dữ liệu mới mà chúng tôi sử dụng để đào tạo mô hình. Bộ dữ liệu phần lớn là một tập hợp các hình ảnh được liên kết với một tập hợp được xác định trước của các từ; tức là, một vốn từ vựng của các từ. Chúng tôi sử dụng tập dữ liệu này cho các nhiệm vụ được mô tả trong Chương 4 yêu cầu phải nắm được vốn từ vựng của tất cả các nhiệm vụ.

Tuổi của sự tiếp thu

Phần lớn vốn từ vựng của chúng tôi đến từ ngữ liệu Age of Acquisition (AoA) [22], đây là danh sách 30.121 từ tiếng Anh có xếp hạng độ tuổi của từng từ trong số những từ này

đã học.1 Các từ được học ở độ tuổi trẻ hơn có xu hướng cụ thể hơn [41], [4], và trong các từ cụ thể có xu hướng trở nên thực tế hơn [43]. Điều này được hỗ trợ bởi AoA tập dữ liệu–để trực quan hóa điều này, chúng tôi đã kết hợp tập dữ liệu AoA với tập dữ liệu xếp hạng tính cụ thể2 trong đó các từ được xếp hạng theo thang điểm từ 1 đến 5 (1 là mức thấp nhất cụ thể và 5 là cụ thể nhất) và vẽ đồ thị mức độ cụ thể trung bình của các từ được học từ độ tuổi 1-19, được thể hiện trong Hình 3.1. Bộ dữ liệu này cung cấp cho chúng ta một vốn từ vựng lớn điểm bắt đầu chứa những từ có khả năng cụ thể hơn từ vựng được lấy từ các tập dữ liệu khác. Hình 3.2 cho thấy sự phân bố đều đặn của bê tông và những từ trừu tượng, nghĩa là cả những từ trừu tượng và cụ thể sẽ được thể hiện trong mô hình cuối cùng đã được đào tạo.

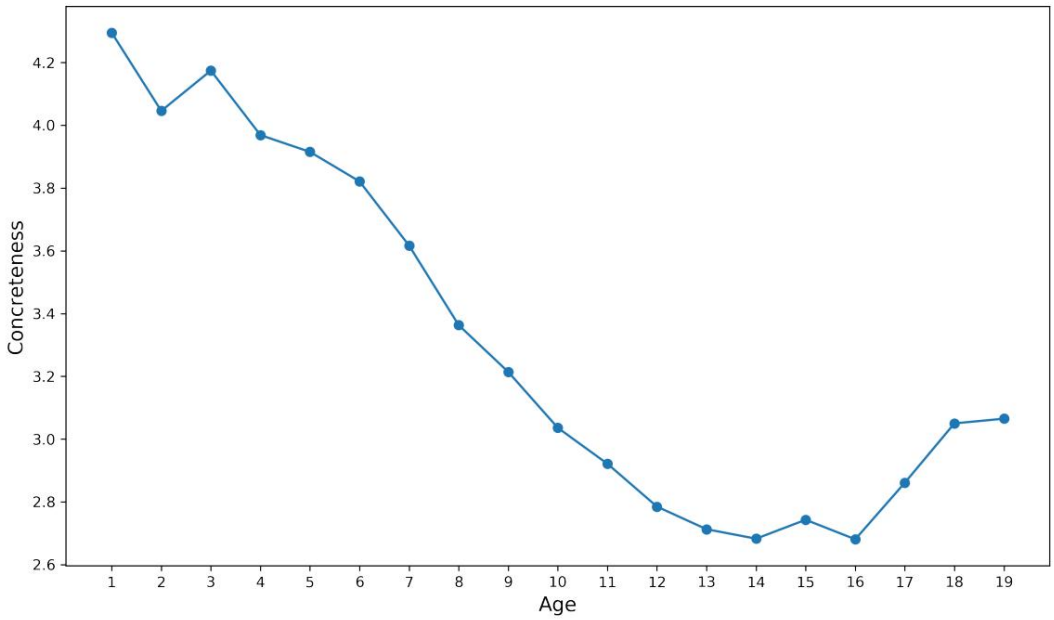
CoNLL-2000

Bộ dữ liệu này [37] chứa khoảng 10.948 câu tiếng Anh và chỉ định một số các loại cụm từ, bao gồm cụm danh từ, cụm động từ và cụm tính từ, trong số những người khác. Mỗi từ trong tập dữ liệu CoNLL-2000 được chú thích bằng một phần của bài phát biểu (POS) thẻ và thẻ chỉ ra cả loại khối và liệu đó có phải là khối đầu tiên hay không token của khối. Chúng tôi sử dụng điều này trong nhiệm vụ phân cụm cụm từ để đánh giá mô hình của chúng tôi, được giải thích bên dưới. Bộ dữ liệu CoNLL-2000 có vốn từ vựng là 18.142 từ.

CoNLL-2003

Ngữ liệu CoNLL-2003 [38] bao gồm khoảng 22.137 câu tiếng Anh và chỉ rõ bốn loại thực thể được đặt tên: địa điểm, tổ chức, cá nhân và các loại khác (ví dụ các thực thể không phù hợp với bất kỳ loại nào trong ba loại đã đề cập ở trên). Mỗi từ trong

1http://crr.ugent.be/archives/806
2http://crr.ugent.be/archives/1330



Hình 3.1: Tương quan giữa tính cụ thể của từ và độ tuổi học những từ đó. Như có thể thấy, những từ học được ở độ tuổi nhỏ hơn thì rất cụ thể, còn những từ học được ở độ tuổi lớn hơn thì trừu tượng hơn. Biểu đồ bắt đầu có xu hướng tăng lên ở độ tuổi cao nhất, đặc biệt là ở độ tuổi 18 và 19, mặc dù cần lưu ý rằng ở độ tuổi này, chỉ có 5 và 2 từ tương ứng trong tập dữ liệu AoA.

tập dữ liệu được chú thích bằng thẻ POS, thẻ đoạn cú pháp và thẻ thực thể được đặt tên

điều đó chỉ rõ liệu từ đó có phải là một thực thể được đặt tên hay không và nếu đó là một thực thể được đặt tên thì nó thuộc loại nào

là. Từ vựng của tập dữ liệu CoNLL-2003 là 23.625 từ.

WordSim-353

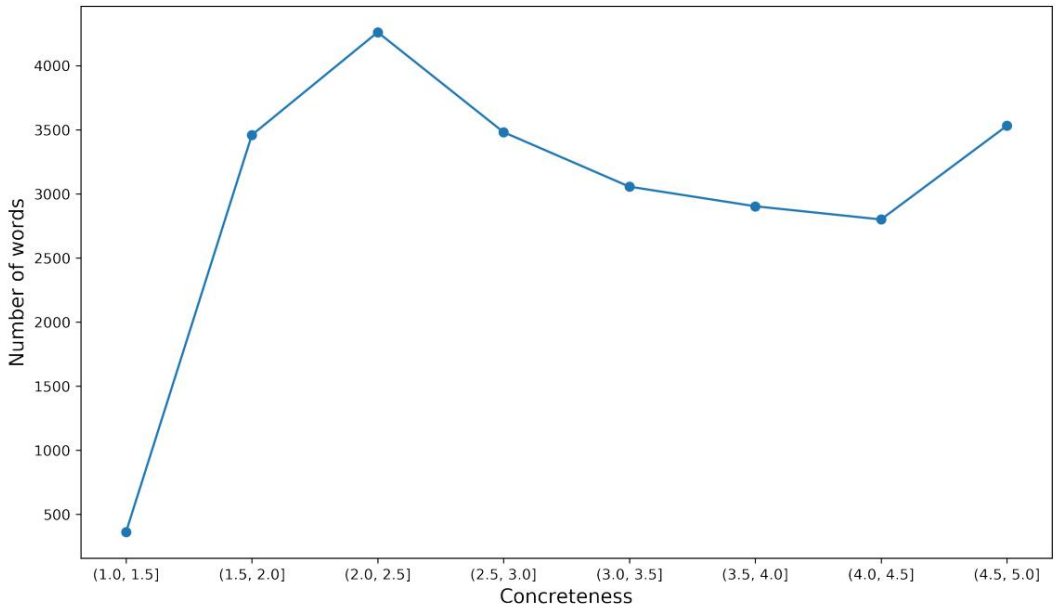
Chúng tôi sử dụng tập dữ liệu WordSim-353 [12]–cụ thể là tập dữ liệu tương tự WordSim-353

được cung cấp bởi các tác giả trong [2]. Bộ dữ liệu này bao gồm 203 cặp danh từ, mỗi cặp có một

điểm tương đồng, mà chúng tôi sử dụng trong nhiệm vụ tương đồng từ. WordSim-353 ban đầu

tập dữ liệu kết hợp các cặp tương đồng (ví dụ: “ô tô” và “xe tải”) với các cặp liên quan (ví dụ:

“xe hơi” và “đường bộ”); chúng tôi sử dụng phiên bản được cung cấp trong [2] vì nó chia tách WordSim-353



Hình 3.2: Trong biểu đồ này, các từ được đặt vào các thùng cụ thể tăng dần 0,5 (ví dụ: (2,5, 3,0], là một phạm vi bao gồm các giá trị 2,5 và không bao gồm các giá trị 3,0). Có sự phân bố khá đồng đều các mức độ cụ thể trong tập dữ liệu AoA, ngoại trừ phạm vi cụ thể (1,0, 1,5] có ít từ.

vào các tập dữ liệu liên quan và tương đồng riêng biệt. Tập dữ liệu tương đồng WordSim-353 có vốn từ vựng là 277 từ.

SimLex-999

Chúng tôi cũng kiểm tra nhiệm vụ tương tự từ trên tập dữ liệu SimLex-999 [15], điều này mới hơn WordSim-353 và chứa 999 cặp từ: 666 cặp danh từ, 222 cặp động từ, và 111 cặp tính từ. Bộ dữ liệu này đặc biệt đề cập đến vấn đề tương đồng và sự liên quan được tìm thấy trong WordSim-353 và do đó chỉ chứa các cặp tương đồng. Giống như tập dữ liệu WordSim-353, SimLex-999 bao gồm điểm tương đồng cho mỗi cặp từ, cùng với thẻ POS, điểm liên kết (tức là liên quan) và xếp hạng tính cụ thể cho mỗi từ. Bộ dữ liệu SimLex-999 có vốn từ vựng là 1.028 từ.

Quay số

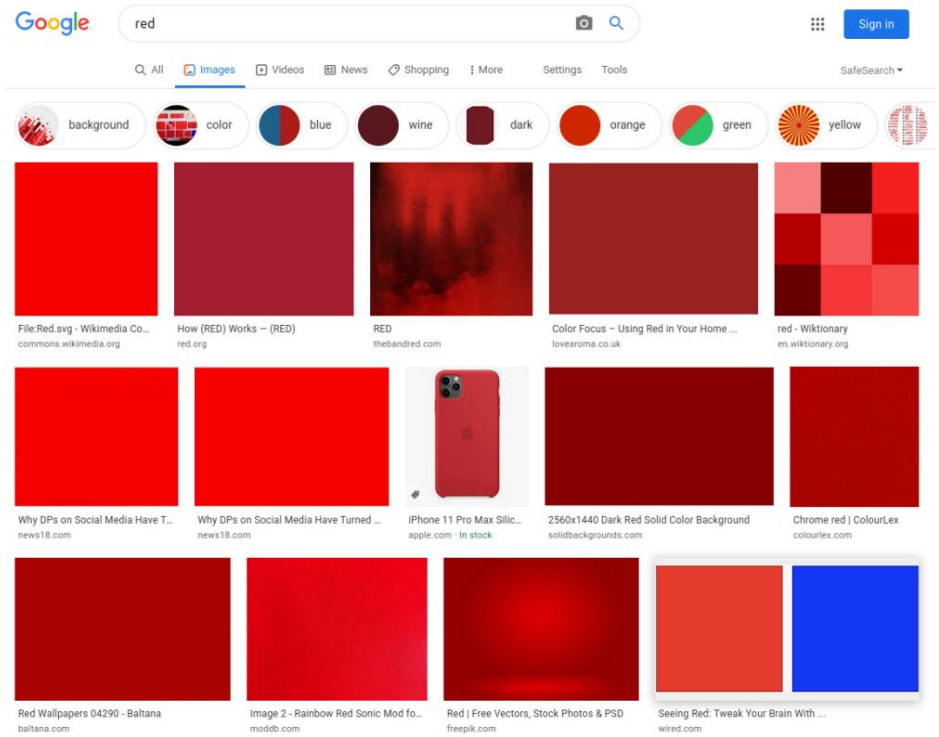
Bộ dữ liệu VisDial [10] bao gồm cả hình ảnh và văn bản và về cơ bản là một bộ sưu tập
của các câu hỏi và câu trả lời về hình ảnh. Chúng tôi sử dụng cả hai phiên bản 0.9 và 1.0 của
tập dữ liệu trong một nhiệm vụ đối thoại trực quan. Phần hình ảnh (mà chúng tôi không sử dụng trực tiếp)
bao gồm khoảng 120.000 hình ảnh từ “Microsoft Common Objects in Context”
(MSCOCO) tập dữ liệu [25], với việc bổ sung khoảng 10.000 hình ảnh từ Flickr trong
VisDial v1.0.3 Phần còn lại của tập dữ liệu VisDial là tập hợp 10 vòng
đối thoại cho mỗi hình ảnh được đưa ra từ tập dữ liệu, cùng với danh sách tất cả các câu hỏi
và danh sách tất cả các câu trả lời. Các vòng đối thoại này bao gồm ID hình ảnh, chú thích
của hình ảnh, ID của câu hỏi về hình ảnh, 100 ID câu trả lời có thể có và
ID của câu trả lời đúng. Đối với cả hai phiên bản của tập dữ liệu, chúng tôi đã đào tạo bằng cách sử dụng
bộ đào tạo và được đánh giá trên bộ xác thực, tất cả đều được cung cấp bởi
người duy trì tập dữ liệu VisDial. Tập dữ liệu VisDial v0.9 có 9.697 từ trong
từ vựng, trong khi tập dữ liệu v1.0 có 11.166 từ.

Bộ dữ liệu cuối cùng Sau khi kết hợp danh sách từ AoA ban đầu với danh sách từ bổ sung
từ các tập dữ liệu được mô tả ở trên, chúng tôi đã đưa ra tổng cộng 61.003 từ trong
từ vựng.

3.1.2 Thu thập dữ liệu hình ảnh

Để đào tạo mô hình của chúng tôi, chúng tôi cần dữ liệu được chú thích. Sau [19], chúng tôi bắt đầu với
từ vựng của chúng tôi, sau đó đối với mỗi thuật ngữ, chúng tôi thực hiện tìm kiếm hình ảnh trên Google cho thuật ngữ đó

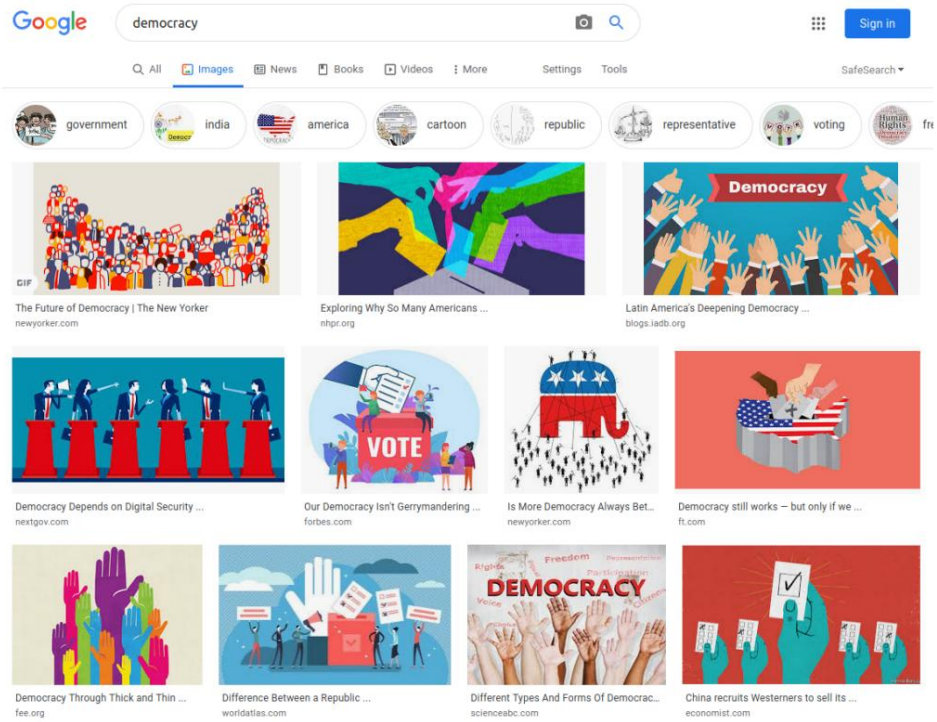
3Mô tả chi tiết hơn về tập dữ liệu có thể được tìm thấy tại https://visualdialog.org/challenge/2020#dataset-description



Hình 3.3: Kết quả tìm kiếm hình ảnh của Google cho từ “đỏ”.

và lấy 100 hình ảnh hàng đầu cho thuật ngữ đó. Chúng tôi quyết định 100 hình ảnh cho mỗi từ vì nó có kích thước hợp lý mà không quá lớn (chẳng hạn như 1000 hình ảnh cho mỗi từ), mặc dù [17] đã chứng minh rằng các bộ phân loại WAC có thể được đào tạo hiệu quả trên ít nhất 10 hình ảnh. Việc tìm kiếm và lưu những hình ảnh này theo cách thủ công sẽ mất rất nhiều thời gian thời gian, vì vậy chúng tôi đã sử dụng một tập lệnh để tự động hóa quá trình này. Do những thay đổi trong Google trang tìm kiếm hình ảnh, tập lệnh ngừng hoạt động giữa chừng khi đang tải xuống hình ảnh cho tập dữ liệu của chúng tôi và chúng tôi phải thay đổi tập lệnh để nó tải xuống hình ảnh từ Bing thay vào đó (khoảng 500 từ hoặc 5.000 hình ảnh).

Mục đích là mỗi hình ảnh này đại diện cho từ được sử dụng trong tìm kiếm-ví dụ, tìm kiếm “màu đỏ” sẽ cho kết quả là hình ảnh có chứa màu đó màu đỏ trong đó, như được thấy trong Hình 3.3. Điều này hiệu quả hơn đối với các từ cụ thể (như màu sắc



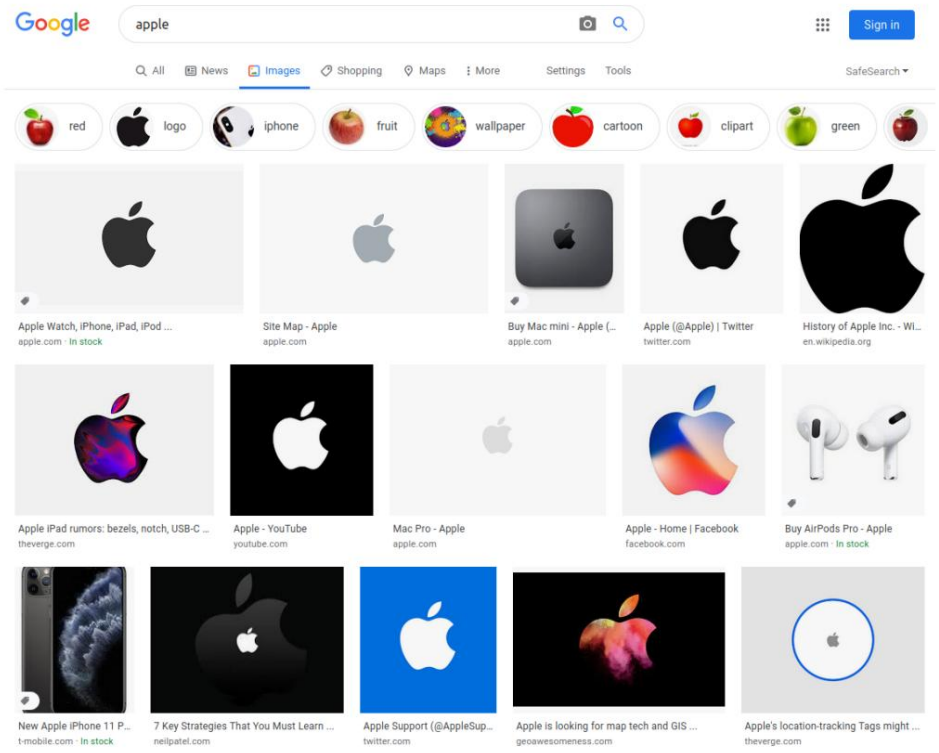
Hình 3.4: Kết quả tìm kiếm hình ảnh của Google cho từ “dân chủ”.

từ) hơn là những từ trừu tượng hơn, như dân chủ, được thể hiện trong Hình 3.4–xem xét
Nhìn những hình ảnh này, người ta có thể nghĩ rằng dân chủ là một đám đông bàn tay giơ lên cao.

Một hạn chế của tập dữ liệu này là nó không cung cấp sự giải thích rõ ràng về ý nghĩa. Đối với
Ví dụ, từ “táo” có thể ám chỉ cả trái cây (một vật thể vật lý, có mặt đất)
và công ty công nghệ. Sự thiếu hụt sự mơ hồ về ý nghĩa này đưa tiếng ồn vào
tập dữ liệu; như có thể thấy trong 3.5, tìm kiếm hình ảnh của Google chỉ hiển thị hình ảnh của Apple
logo cho tìm kiếm này. Chúng tôi để lại sự mơ hồ về ý nghĩa cho công việc trong tương lai.

3.2 WAC

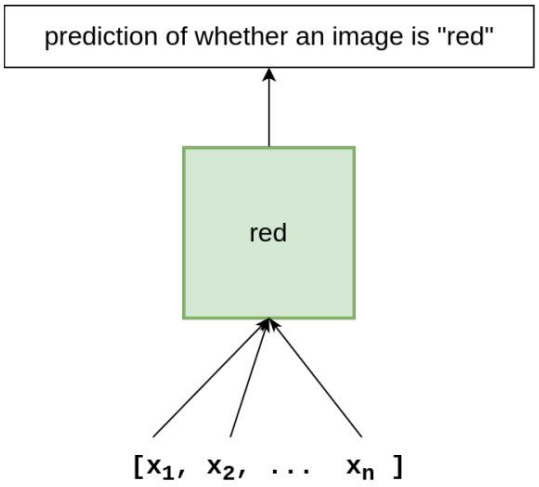
WAC được giới thiệu trong [17]. Tiếp theo [39], phương pháp WAC đối với ngữ nghĩa từ vựng
về cơ bản là một cách tiếp cận độc lập với nhiệm vụ để dự đoán tính phù hợp về mặt ngữ nghĩa của



Hình 3.5: Kết quả tìm kiếm hình ảnh của Google cho từ “apple”.

từ trong ngữ cảnh vật lý. Mô hình WAC ghép nối mỗi từ w trong vốn từ vựng V của nó
với một bộ phân loại ánh xạ các đặc điểm có giá trị thực của một đối tượng thành một ngữ nghĩa thích hợp-
điểm ness.

Ví dụ, để tìm hiểu ý nghĩa cơ bản của từ màu đỏ, hình ảnh cấp thấp
các đặc điểm của tất cả các hình ảnh được mô tả bằng từ đỏ trong một tập hợp các chú thích hình ảnh hoặc
các mô tả được đưa ra như các trường hợp tích cực cho bộ phân loại học có giám sát (ví dụ:
hồi quy logistic hoặc perceptron nhiều lớp), được mô tả trong Hình 3.6. Những hình ảnh này
các tính năng có thể là, ví dụ, các giá trị RGB của một hình ảnh hoặc bất kỳ lớp trên cùng nào của
một CNN (tức là, chuyển giao học tập), như trong [18], chúng tôi theo dõi ở đây. Như mỗi phân loại
là một bộ phân loại nhị phân, nó cần các trường hợp dương và âm; 3 trường hợp âm
được lấy mẫu ngẫu nhiên từ tập hợp hình ảnh bổ sung (tức là hình ảnh có



Hình 3.6: Cấu trúc của mô hình WAC cho từ red. Các đặc điểm của một tập hợp các hình ảnh được mô tả bởi từ red được truyền làm đầu vào để đào tạo một bộ phân loại nhị phân (như hồi quy logistic). Sau khi đào tạo, bộ phân loại có thể trả về dự đoán về việc một hình ảnh nhất định có thuộc lớp ngữ nghĩa “red” hay không.

chú thích không chứa từ đỏ). Điều này dẫn đến một bộ phân loại được đào tạo, mà các đặc điểm của hình ảnh có thể được áp dụng để xác định mức độ nhận dạng hình ảnh đó như màu đỏ; nói cách khác, dự đoán xem hình ảnh có màu “đỏ” hay không.

3.3 Nhúng WAC

Một sản phẩm phụ quan trọng của phương pháp WAC là mỗi từ tạo ra một cá nhân phân loại và mỗi phân loại có cấu trúc bên trong mà chúng ta có thể sử dụng cho các mục đích khác ngoài phân loại nhị phân. Hầu hết các bộ phân loại (ví dụ, hồi quy logistic, nhiều lớp perceptron) được mô hình hóa bằng cách sử dụng một tập hợp các hệ số, thường có một hệ số cho mỗi tính năng đầu vào. Chúng ta có thể xử lý các hệ số này như một vectơ và sử dụng chúng trong một không gian nhúng, như được thực hiện trong các phương pháp tiếp cận ngữ nghĩa phân phối.4 Một nhúng sẽ có thông tin về các khía cạnh cơ sở trực quan của một từ,

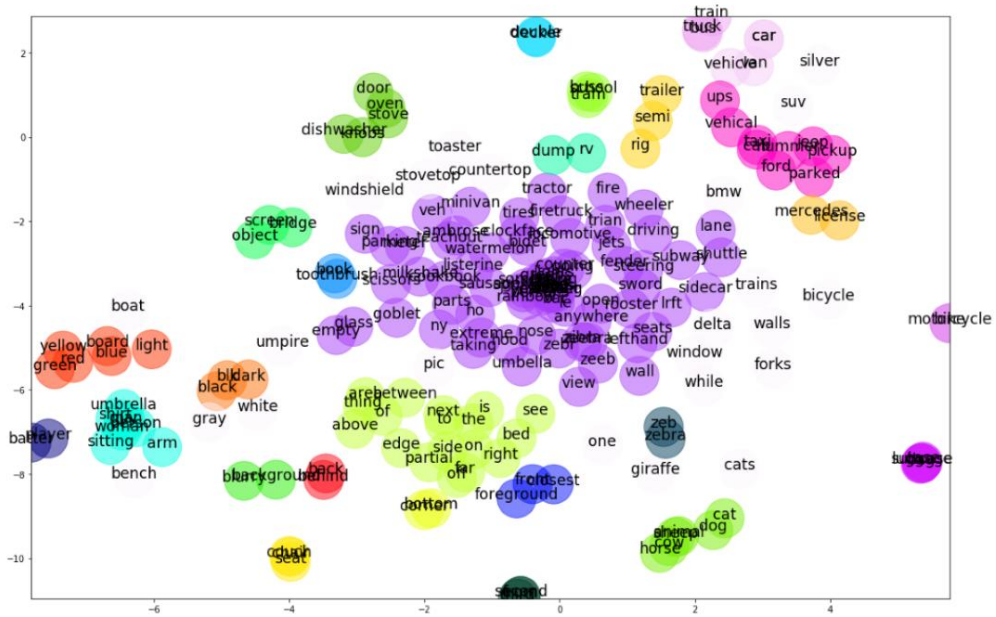
4Phương pháp này ban đầu được giới thiệu với Word2vec [32].

đó là thông tin mà nhúng tiêu chuẩn không có vì chúng chỉ
được đào tạo trên văn bản. Hơn nữa, việc nhúng như vậy có thể dễ dàng kết hợp với các
nhúng hiện có (ví dụ, thông qua nối).

Công trình trước đây của chúng tôi trong Moro et al. đã khám phá tính khả thi của việc sử dụng các vectơ này như
nhúng từ [33], sử dụng kiến trúc mạng nơ-ron đơn giản của một lớp ẩn
chứa 3 nút. Sau [32], các bộ phân loại WAC đơn giản đã được đào tạo bằng cách sử dụng các từ
và hình ảnh trong tập dữ liệu MSCOCO và các nhúng hệ số đã được trích xuất
để xem liệu chúng có tập hợp theo cách có thể mong đợi trong nhúng phân phối hay không.
Hình 3.7 cho thấy kết quả của việc ánh xạ các hệ số vào 2 chiều bằng cách sử dụng t-
Phân phối Nhúng hàng xóm ngẫu nhiên (TSNE) [30] và nhóm các kết quả
với Phân cụm không gian dựa trên mật độ của các ứng dụng có tiếng ồn (dbscan) [40]. Từ
trong con số này, một số cụm đáng chú ý bao gồm:

- vàng, đỏ, xanh lá cây, xanh dương, ánh sáng, bảng
- diện tích, giữa, của, phía trên, cạnh, bên cạnh, bên phải
- mèo, chó, ngựa, bò, cừu, động vật

Những kết quả này cho thấy rằng các hệ số phân loại có thể được sử dụng như các vectơ cho
nhúng từ, vì chúng thể hiện sự tương đồng với nhúng, như nhúng cho
các từ tương tự cũng có xu hướng gần nhau hơn trong không gian vectơ; tức là khoảng cách cosin
giữa hai từ thể hiện mức độ giống nhau về mặt ngữ nghĩa của chúng. Chúng tôi đánh giá thêm
và cải thiện mô hình của Moro et al. [33], và chúng tôi gọi biểu diễn vectơ của chúng tôi là
bộ phân loại có cơ sở wac2vec.

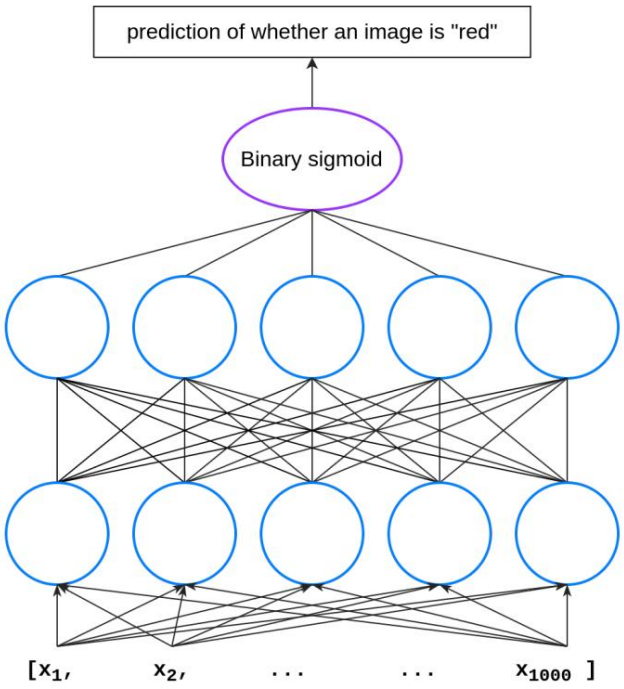


Hình 3.7: Các cụm hệ số lớp ẩn của bộ phân loại, như được thể hiện trong Moro et al.[33]. Từ vựng là 100 từ hàng đầu trong tập dữ liệu MSCOCO.

3.4 Đào tạo wac2vec

Các thử nghiệm ban đầu sử dụng phương pháp của Moro et al. [33] cho thấy kết quả hỗn hợp về mặt ngữ nghĩa nhiệm vụ tương tự; trong luận án này chúng tôi sử dụng một kiến trúc có nguyên tắc hơn, một kiến trúc mới hơn CNN để trích xuất các đặc điểm hình ảnh và giảm nhúng bằng PCA để loại bỏ tiếng ồn.

Chúng tôi đã đào tạo các bộ phân loại WAC trên tập dữ liệu hình ảnh của Google được mô tả trong Chương 3.1. Để có được các đặc điểm hình ảnh, chúng tôi đã truyền hình ảnh qua EfficientNet, một mô hình CNN mới được Google phát hành gần đây đạt được hiệu suất tiên tiến trên một số của các tập dữ liệu [42]. Điều này dẫn đến một biểu diễn vectơ của hình ảnh có 1000 kích thước (tức là chúng tôi sử dụng lớp ngay bên dưới lớp dự đoán; chúng tôi đã xác định lớp này thực hiện tốt nhất thông qua một tập hợp con dữ liệu của chúng tôi về tác vụ VisDial). 1000 vectơ đặc trưng cho mỗi từ được sử dụng làm đặc trưng đầu vào cho WAC tương ứng



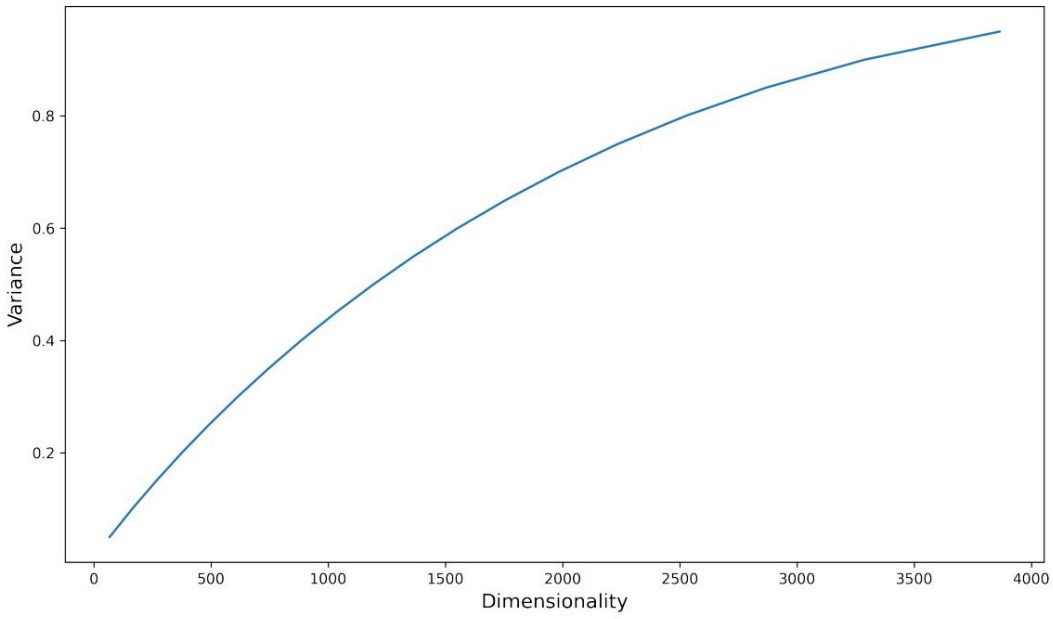
Hình 3.8: Cấu trúc của bộ phân loại wac2vec red, được đào tạo trên 100 hình ảnh mô tả từ red. Bộ phân loại này lấy đầu vào là 1000 đặc điểm, có hai lớp ẩn, mỗi lớp có 5 nút và một sigmoid nhị phân đưa ra dự đoán về việc liệu một hình ảnh nhất định có thuộc lớp ngữ nghĩa “red” hay không. Lớp dưới cùng có 5005 hệ số (5000 trọng số + 5 thuật ngữ bias), lớp trên cùng có 30 hệ số (25 trọng số + 5 thuật ngữ bias) và sigmoid nhị phân cuối cùng có 6 hệ số (5 trọng số + 1 thuật ngữ bias).

bộ phân loại, cũng như các vectơ đặc trưng cho 3 từ được lấy mẫu ngẫu nhiên, được dùng như những ví dụ đào tạo tiêu cực.

Theo [33], mô hình WAC của chúng tôi là một perceptron nhiều lớp. Các bộ phân loại của chúng tôi có kiến trúc sau: Hai lớp ẩn, mỗi lớp bao gồm 5 nút; chúng tôi đã xác định theo kinh nghiệm, kiến trúc này hoạt động tốt nhất khi sử dụng một tập hợp con dữ liệu của chúng tôi trên một phần của tác vụ VisDial. Các lớp ẩn sử dụng hàm kích hoạt tanh, hàm này bảo toàn cực tính của các hệ số (tức là, giá trị hệ số nằm trong khoảng từ -1 đến 1), với nhị phân lớp trên cùng sigmoid. Chúng tôi đã đào tạo bằng cách sử dụng bộ giải adam (giá trị alpha là 0,1 được xác định

thông qua thử nghiệm), một thuật toán tối ưu hóa hiệu quả hội tụ nhanh chóng. Tốt nhất
các tham số đã được tìm thấy bằng cách sử dụng một tập hợp con của dữ liệu đào tạo. Các bộ phân loại WAC là
được đào tạo trong 500 kỷ nguyên tối đa, với cơ chế dừng sớm của scikit-learn
dừng tập luyện khi tình trạng mất mát không còn được cải thiện.5

Sau khi đào tạo từng bộ phân loại trên các hình ảnh liên quan của nó, chúng tôi đã lấy các hệ số từ
cả hai lớp ẩn (kiểm tra cho thấy việc lấy tất cả các hệ số hoạt động tốt nhất),
tạo ra các vectơ có chiều là 5041 (tức là 5 nơ-ron ở lớp thấp nhất,
mỗi lớp có 1000 tính năng đầu vào; 5 nơ-ron ở lớp giữa với 5 đầu vào mỗi lớp từ
các nút trước đó; 5 đầu vào cho lớp sigmoid cuối cùng; và các điều khoản độ lệch cho mỗi nút).



Hình 3.9: Đồ thị về chiều kết quả sau khi giảm các vectơ wac2vec thành các phương sai khác biệt.

Các vectơ này lớn hơn nhiều so với hầu hết các nhúng có xu hướng và thử nghiệm ban đầu
cho thấy rằng họ thực hiện kém nhiệm vụ đối thoại trực quan, có thể là do tiếng ồn

5Ban đầu chúng tôi sử dụng Keras, một thư viện mạng nơ-ron, nhưng việc tuần tự hóa và hủy tuần tự hóa các dữ liệu đã được đào tạo mô hình rất chậm, vì vậy chúng tôi chuyển sang scikit-learn, nhanh hơn nhiều.

trong các vectơ. Sử dụng PCA, chúng tôi đã giảm chiều của các vectơ và chọn
để cân bằng giữa phương sai và kích thước vectơ. Hình 3.9 cho thấy biểu đồ của
phương sai so với chiều. Dựa trên điều này, chúng tôi đã chọn chiều là 1700,
vì đó là nơi chúng tôi tin rằng có sự cân bằng tốt nhất giữa phương sai và tính đa chiều.
Tuy nhiên, chúng tôi đã hết bộ nhớ GPU khi sử dụng chiều này trên hình ảnh
nhiệm vụ đối thoại, vì vậy chúng tôi đã giảm chiều không gian xuống 100 cho đến khi nhiệm vụ có thể chạy
thành công với các vectơ wac2vec, cuối cùng đạt được số chiều là 1300.

Chúng tôi đã lấy các vectơ wac2vec 1300 chiều thu được và đánh giá hiệu suất của chúng
mance trong các nhiệm vụ NLP được mô tả trong Chương 4: sự giống nhau của từ, thực thể được đặt tên
nhận dạng, phân cụm cụm từ và đối thoại trực quan. Khi nhiệm vụ đối thoại trực quan sử dụng
dữ liệu văn bản được dựa trên hình ảnh, chúng tôi sử dụng nó làm tiêu chuẩn để cải thiện thử nghiệm
trong sự kết hợp của wac2vec và nhúng.

Giả định Mô hình wac2vec của chúng tôi đưa ra hai giả định quan trọng:

1. ngữ nghĩa từ vựng của một từ là độc lập với tất cả các từ khác

2. tất cả các từ đều có nguồn gốc vật lý trong thế giới vật chất

Đối với Giả định 1, điều này có nghĩa là ý nghĩa của một từ không liên quan gì đến
với những từ khác. Rõ ràng là không phải như vậy, vì chúng ta sử dụng các từ theo trình tự, không phải bằng
chính họ. Giả định 2 cũng không đúng trong thế giới thực, vì rõ ràng có những từ
trừu tượng, như “dân chủ”. Như đã thảo luận trước đó trong Phần 3.1, tìm kiếm hình ảnh
kết quả cho nền dân chủ có thể liên tục hiển thị hình ảnh có bàn tay giơ lên,
nhưng điều này không nắm bắt được tất cả những gì mà dân chủ có nghĩa. Những giả định này hướng dẫn
cách chúng tôi thực hiện đánh giá của mình - chúng tôi nối các nhúng wac2vec có cơ sở

với các nhúng được đào tạo bằng cách sử dụng phương pháp phân phối, tạo ra điều ngược lại giả định: rằng nghĩa của một từ phụ thuộc vào các từ khác và rằng tất cả từ ngữ là trừu tượng. Bằng cách kết hợp wac2vec với nhúng phân phối, chúng tôi hướng đến để giảm thiểu những giả định này. Đánh giá mô hình của chúng tôi được giải thích chi tiết trong chương tiếp theo.

CHƯƠNG 4

SỰ ĐÁNH GIÁ

Để thiết lập một đường ống để đánh giá hiệu quả của wac2vec và khám phá những gì nó học hỏi, và hơn nữa để thiết lập các đường cơ sở cho những đánh giá đó, chúng tôi đã tiến hành loạt thí nghiệm cho 4 nhiệm vụ NLP:

- đối thoại trực quan
- cụm từ phân đoạn
- nhận dạng thực thể được đặt tên
- sự tương đồng về mặt ngữ nghĩa

Chúng tôi đã sử dụng nhiệm vụ đối thoại trực quan làm nhiệm vụ chính để thử nghiệm các cải tiến mô hình của chúng tôi. Các nhiệm vụ còn lại được chọn vì chúng đại diện cho các nhiệm vụ theo thứ tự phức tạp, bắt đầu với các nhiệm vụ thường được sử dụng trước khi được áp dụng trong các nhiệm vụ: sự tương đồng về mặt ngữ nghĩa, nhận dạng thực thể được đặt tên và phân cụm cụm từ. Những nhiệm vụ cung cấp cái nhìn sâu sắc về những gì mô hình của chúng tôi học được (hoặc không học được). Là một nhiệm vụ với dữ liệu có căn cứ, bất kỳ cải tiến nào về hiệu suất khi kết hợp wac2vec với nhúng truyền thống có nhiều khả năng được hiển thị trong nhiệm vụ đối thoại trực quan. Trong nhiệm vụ tiếp theo, chúng tôi kiểm tra khả năng hiểu cú pháp của wac2vec bằng cách phân cụm cụm từ và các nhiệm vụ nhận dạng thực thể được đặt tên, và cuối cùng thực hiện một nhiệm vụ đơn giản, tính tương đồng của từ, như một thử nghiệm về cụm ngữ nghĩa của các nhúng wac2vec có cơ sở. Trong mỗi

những nhiệm vụ này, chúng tôi đã mã hóa dữ liệu đầu vào văn bản bằng 1) các mô hình nhúng từ chuẩn, bao gồm GloVe, fastText và BERT, 2) hệ số wac2vec và 3) nhúng được nối với hệ số wac2vec. Sau đây là mô tả ngắn gọn về NLP các tác vụ mà chúng tôi đã áp dụng wac2vec, bao gồm ví dụ về từng tác vụ.

Như đã đề cập ở trên, ngoài GloVe, chúng tôi cũng chạy các tác vụ này bằng BERT như nhúng phân phối. BERT là một kiến trúc mạng nơ-ron gần đây đã thiết lập trạng thái nghệ thuật cho nhiều nhiệm vụ NLP. Chúng tôi đặt một từ duy nhất tại một thời điểm thông qua BERT để tạo nhúng BERT; trong các tác vụ như trực quan đối thoại sử dụng một câu đầy đủ, chúng ta muốn chuyển toàn bộ câu thành BERT để sử dụng ngữ cảnh (tức là “đứng trên bờ sông” so với “cô ấy đã cướp ngân hàng”). Tuy nhiên, do hạn chế về phần cứng, chúng tôi không thể sử dụng BERT theo cách này cho nhiệm vụ đối thoại trực quan, mặc dù chúng ta có thể làm như vậy trong nhiệm vụ nhận dạng thực thể được đặt tên và phân nhóm cụm danh từ.

4.1 Đối thoại trực quan

Không giống như những nhiệm vụ khác, nhiệm vụ này kết hợp yếu tố trực quan ngoài dữ liệu văn bản. Trong đối thoại trực quan, mục tiêu là trả lời các câu hỏi về hình ảnh, đặc biệt là trong bối cảnh của các câu hỏi đã hỏi và trả lời trước đó. Nói cách khác, nhiệm vụ này kết hợp cả ngữ cảnh trực quan và ngữ cảnh từ vựng.

Câu hỏi: Có bao nhiêu người?

[Trả lời: Hai]

Câu hỏi: Họ có đứng không?

[Trả lời: Không, họ đang ngồi]

4.1.1 Nhiệm vụ & Thủ tục

Chúng tôi làm theo Massiceti et al. về vấn đề này, bằng cách sử dụng mô hình được tạo ra và cung cấp bởi [31], sử dụng tập dữ liệu MSCOCO [25] và chúng tôi đã sử dụng cả GloVe và BERT nhúng-dings.1 Ngoài việc chỉ định chiều cuối cùng của các nhúng được cung cấp (tức là wac2vec được nối với BERT) và đặt kích thước lô thành 16 (theo thứ tự để tránh hết bộ nhớ), chúng tôi giữ nguyên tất cả các thiết lập như đã được xác định trong Massiceti et al. Trong cách tiếp cận của họ, bối cảnh đối thoại trước đó và các đặc điểm của hình ảnh đang được thảo luận không được sử dụng; các tác giả khẳng định rằng nhiệm vụ có thể được tiếp cận chỉ với cặp câu hỏi-câu trả lời và khi các vectơ wac2vec được áp dụng trực tiếp vào các từ trong các cặp câu hỏi-trả lời, chúng ta không cần hình ảnh để đánh giá hiệu quả của mô hình của chúng tôi trong nhiệm vụ này. Mô hình của họ học cách nhúng chung giữa các câu hỏi và trả lời bằng cách tính toán ma trận chiếu (một cho câu hỏi và một cho câu trả lời), với mục tiêu tối đa hóa mối tương quan giữa các phép chiếu của mỗi ma trận. Tại thời gian kiểm tra, câu trả lời của ứng viên được xếp hạng theo khoảng cách cosin giữa khớp

1Mã chúng tôi sử dụng cho nhiệm vụ này có thể được tìm thấy tại https://github.com/danielamassiceti/CCA-visualdialogue

Bảng 4.1: Kết quả từ nhiệm vụ đối thoại trực quan. Nửa trên của kết quả dành cho tập dữ liệu v0.9 và nửa dưới dành cho tập dữ liệu v1.0. Từ trái sang phải, các số liệu là: thứ hạng trung bình; khả năng nhớ lại ở mức 1, 5 và 10; và thứ hạng trung bình đối ứng. Kết quả tốt nhất được in đậm. Đối với kết quả tốt nhất kết hợp wac2vec và nhúng, chúng tôi đã tính toán ý nghĩa thống kê bằng cách sử dụng kiểm định t ghép cặp, với alpha là 0,05. Giả thuyết vô hiệu của chúng tôi là wac2vec không cải thiện hiệu suất của nhúng phân phối. Với giá trị p <0,01 cho mỗi kết quả mà chúng tôi đã kiểm tra, chúng tôi thấy rằng những kết quả có hiệu suất tốt nhất này có ý nghĩa thống kê.

Đường	Ông R@1		R@5	R@10 MRR	
cơ sở của mô hình (fastText) 16.2052 16.8566 44.9837 58.0817 0.3043					
Găng tay 18.6441 13.9362 38.1933 51.7285 0.2623					
BERT 14.4935 18.4362 47.0531 60.9851 0.3530					
wac2vec 15.3334 19.1011 46.7808 60.9582 0.3249					
wac2vec + GloVe 14.9333 20.3313 49.2361 62.6997 0.3409					
wac2vec + BERT 14.9250 20.7824 49.2855 62.3896 0.3441					
Đường cơ sở (fastText) 17,0314 16,0320 41,1822 55,1938 0,2860					
Găng tay 19.9415 13.9244 35.8527 49.6657 0.2540					
BERT 15.5700 17.6744 43.9874 58.2219 0.3052					
wac2vec 15.9998 17.3934 42.9264 58.3333 0.3017					
wac2vec + GloVe 15.4788 18.2897 44.7384 59.6415 0.3131					
wac2vec + BERT 15.6268 18.7888 44.9128 59.1667 0.3166					

nhúng của câu hỏi và câu trả lời của mỗi ứng viên, dựa trên ngữ nghĩa sự giống nhau của các câu hỏi và câu trả lời tiềm năng, thay vì dữ liệu hình ảnh. Chúng mã hóa câu hỏi và câu trả lời sử dụng nhúng fastText; chúng tôi thay thế chúng bằng của riêng chúng tôi (wac2vec, GloVe, BERT, wac2vec + GloVe và wac2vec + BERT).

Chúng tôi đã thử nghiệm trên cả hai phiên bản của tập dữ liệu mà họ sử dụng (tức là phiên bản 0.9 và 1.0), được cung cấp bởi [10].

4.1.2 Số liệu

Trong nhiệm vụ này, mục tiêu là xếp hạng 100 câu trả lời có thể có cho một câu hỏi về một hình ảnh, chỉ có một câu trả lời là đúng. Nhiệm vụ này sử dụng một số số liệu: 1) thứ hạng trung bình (tức là vị trí xếp hạng trung bình của câu trả lời đúng); 2) nhớ lại r@k

xét đến các vị trí @5, @10 và @15; và 3) thứ hạng tương hỗ trung bình mrr,
trong đó ranki là vị trí của câu trả lời đúng cho câu hỏi thứ i trong Q
tập dữ liệu.

$$r@k = \frac{TP}{TP + FN} \tag{4.1}$$

$$\text{Ông} = \frac{1}{|\text{Hỏi}|}\sum_{\text{tôi}=1}^{|\text{Hỏi}|}\frac{1}{\text{xếp hạng}} \tag{4.2}$$

4.1.3 Kết quả

Bảng 4.1 hiển thị kết quả từ nhiệm vụ đối thoại trực quan, trên cả phiên bản 0.9 và
1.0 của tập dữ liệu đối thoại trực quan. Như có thể thấy, kết hợp wac2vec với GloVe
và nhúng BERT cải thiện hiệu suất trên hầu hết mọi số liệu cho cả hai tập dữ liệu.
Hơn nữa, wac2vec tự nó hoạt động tốt hơn cả GloVe và fastText
đường cơ sở được Massiceti và cộng sự sử dụng trên cả hai phiên bản của tập dữ liệu VisDial.

Những kết quả này cho thấy wac2vec có thể đóng góp thêm vào ngữ nghĩa
sự hình thành của BERT, đã chứng minh hiệu suất tiên tiến trên một
số lượng nhiệm vụ NLP, cũng như GloVe, không có riêng. Đây là một phần
đặc biệt thú vị cho nhiệm vụ này vì cả BERT, GloVe và wac2vec thực sự không
kiểm tra trực quan các hình ảnh trực tiếp–thay vào đó, họ chỉ sử dụng câu hỏi-trả lời
cặp, nhưng wac2vec bổ sung thông tin ngữ nghĩa quan trọng về mặt trực quan. Khi VisDial
tập dữ liệu câu hỏi-trả lời có cơ sở vững chắc (tức là, nhiều từ về những thứ
được thể hiện vật lý trong một bức ảnh), điều đó có nghĩa là wac2vec, một nền tảng
mô hình, sẽ hoạt động tốt ở đây khi được sử dụng như một nhúng cho những từ này. Đây là
được hỗ trợ bởi các kết quả riêng lẻ cho wac2vec, hoạt động tốt hơn

nhúng fastText và GloVe trên tất cả các số liệu.

Hiệu suất kết quả của việc kết hợp wac2vec với nhúng truyền thống trong nhiệm vụ này cho thấy rằng wac2vec cung cấp thông tin ngữ nghĩa nhúng được đào tạo về thiếu văn bản. Tuy nhiên, wac2vec còn học được điều gì nữa? Chúng tôi khám phá điều này trong sau các thí nghiệm.

4.2 Phân cụm cụm từ

Nhiệm vụ phân cụm cụm từ bao gồm việc xác định và trích xuất các cụm từ từ văn bản, và đòi hỏi phải hiểu cú pháp. Ví dụ sau đây mô tả quá trình trích xuất của các cụm danh từ trong một câu.

[Con chó vàng nhỏ] sủa [con mèo].

4.2.1 Nhiệm vụ & Thủ tục

Tiếp theo [3], nhiệm vụ này liên quan đến việc xác định vị trí và phân loại các khối cụm từ trong CoNLL-2000 [37] ngữ liệu, như được mô tả trong Phần 3.1.

Ngoài GloVe và BERT, chúng tôi đã sử dụng nhúng Flair từ thư viện Flair,2 để kiểm tra so với đường cơ sở trong [3]. Thư viện Flair cũng cung cấp nhúng BERT có thể được sử dụng trong nhiệm vụ này, hoạt động trên toàn bộ câu thay vì từng từ riêng lẻ. Theo công việc trong [3], nhiệm vụ này sử dụng LSTM là được đào tạo tối đa 150 kỷ nguyên.3 Các câu được mã hóa bằng cách nhúng và LSTM dự đoán thẻ cụm từ cho mỗi từ trong câu.

2https://github.com/zalandoresearch/flair 3Mã chúng tôi sử dụng cho tác vụ phân cụm cụm từ có sẵn tại https://github.com/flairNLP/flair/blob/master/resources/docs/EXPERIMENTS.md# conll-2000-noun-phrase-chunking-english. Mặc dù trang này gọi đây là phân cụm danh từ, nhưng thực ra đây là tác vụ phân cụm chung và dự đoán nhiều loại phân cụm cụm (ví dụ: cụm động từ và cụm tính từ).

4.2.2 Số liệu

Sau đây [3], kết quả được đánh giá bằng cách sử dụng điểm F1, giá trị trung bình hài hòa của độ chính xác p và nhớ lại r, xác định chính xác cụm từ khóa của các từ trong mỗi câu.

Ngoài ra, chúng tôi cũng báo cáo độ chính xác của việc xác định đúng cụm từ thẻ của mỗi từ. TP, FP, TN và FN là số lượng dương tính thật, dương tính giả, kết quả âm tính thật và kết quả âm tính giả được tìm thấy trong quá trình đánh giá các thẻ cụm từ dự đoán, tương ứng.

$$p = \frac{TP}{TP + FP} \tag{4.3}$$

$$r = \frac{TP}{TP + FN} \tag{4.4}$$

$$F1 = 2 \frac{p \quad r}{p + r} \tag{4,5}$$

$$\text{một} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.6}$$

4.2.3 Kết quả

Trong nhiệm vụ này, BERT từ thư viện Flair đã thực hiện tốt nhất trong số tất cả các nhúng; wac2vec thực hiện thấp nhất. Như thể hiện trong Bảng 4.2, kết hợp wac2vec với các nhúng từ phân phối không mang lại lợi ích gì, ngoại trừ trong sự kết hợp của BERT và wac2vec, mang lại sự cải thiện nhỏ về điểm số F1 và độ chính xác chỉ trên BERT.4 Cụm danh từ là một loại cụm từ được tìm thấy trong

4Chúng tôi không thực hiện các thử nghiệm ý nghĩa thống kê đối với kết quả của thí nghiệm này hoặc các thí nghiệm tiếp theo, vì bất kỳ cải tiến nào khi thêm wac2vec vào nhúng truyền thống đều rất nhỏ và chúng tôi đưa ra

Bảng 4.2: Kết quả từ tác vụ chunking. Hai số liệu ở đây là điểm F1 và độ chính xác. BERT có dấu sao cho biết BERT do thư viện Flair cung cấp. Kết quả tốt nhất được in đậm.

Mô hình	Độ chính xác của điểm F1
Flair (cơ sở)	0,9640 0,9305
wac2vec	0,8550 0,7467
GloVe	0,9382 0,8836
BERT	0,8780 0,7825
BERT*	0,9659 0,9342
wac2vec + Flair 0,9524 wac2vec +	0,9092
GloVe 0,9229 wac2vec + BERT	0,8568
0,8550 wac2vec + BERT* 0,9667	0,7467
0,9356	

tập dữ liệu của nhiệm vụ này và danh từ có xu hướng cụ thể, điều này có thể giải thích tại sao kết quả đã tốt hơn ở đây. Tuy nhiên, vì sự cải thiện là nhỏ, và như wac2vec khác các sự kết hợp không cho thấy sự cải thiện, có thể kết quả này là do ngẫu nhiên.

Kết quả của chúng tôi cho nhiệm vụ này nói chung cho thấy wac2vec không học cú pháp, điều này là điều có thể mong đợi, vì wac2vec đưa ra giả định về tính độc lập của từ vựng.

4.3 Nhận dạng thực thể được đặt tên

Nhận dạng thực thể được đặt tên (NER) liên quan đến việc định vị và phân loại các thực thể được đặt tên trong văn bản. Giống như việc phân cụm cụm từ, nhiệm vụ này cũng chủ yếu là một bài kiểm tra cú pháp. Được đặt tên các thực thể có thể bao gồm con người, tổ chức, địa điểm, cách diễn đạt thời gian và những thứ khác.

[Janet]Person bắt đầu làm việc tại [Twitter]Organization vào năm [2010]

không có kết luận cuối cùng từ họ.

4.3.1 Nhiệm vụ & Thủ tục

Trong nhiệm vụ này, các thực thể được đặt tên đã được xác định và trích xuất từ CoNLL-2003 [38]
ngữ liệu, được mô tả trong Phần 3.1. Như trong nhiệm vụ phân đoạn cụm từ, chúng tôi làm theo [3] và
đã thử nghiệm nhúng Flair ngoài nhúng GloVe và BERT, cũng như
BERT từ thư viện Flair. Cũng giống như nhiệm vụ chunking cụm từ, nhiệm vụ NER sử dụng
một LSTM được đào tạo tối đa 150 kỷ nguyên; đầu vào là các câu được mã hóa
(sử dụng nhúng được cung cấp của chúng tôi) và LSTM dự đoán thẻ thực thể được đặt tên của
mỗi từ trong câu.5

4.3.2 Số liệu

Giống như nhiệm vụ phân cụm cụm từ, sau [3], kết quả được đánh giá bằng cách sử dụng điểm F1
và độ chính xác khi xác định đúng thẻ thực thể có tên của một từ. Hiển thị trong Bảng
4.2 là kết quả của nhiệm vụ này.

4.3.3 Kết quả

Trong nhiệm vụ này, Flair thực hiện tốt hơn wac2vec, GloVe, cả hai phiên bản của BERT,
và tất cả các kết hợp wac2vec + nhúng. Bộ dữ liệu cho nhiệm vụ này, như
như mong đợi, có một số lượng lớn các thực thể được đặt tên như tên và số,
không được thể hiện tốt bằng hình ảnh (tức là chúng là những khái niệm trừu tượng) và
có thể dẫn đến các bộ phân loại WAC kém hiệu quả hơn và do đó các vectơ wac2vec yếu hơn.
Cũng thú vị cần lưu ý là trong khi các kết hợp wac2vec không cung cấp mức cao nhất
kết quả, wac2vec cải thiện kết quả cho nhúng BERT một từ; tuy nhiên,
những kết quả này được thay thế bằng những kết quả của BERT từ thư viện Flair, trong đó BERT

5Đối với nhiệm vụ NER, chúng tôi đã sử dụng mã có sẵn tại https://github.com/flairNLP/flair/blob/
chủ/tài nguyên/tài liệu/TRÍCH DẪN.md#conll-03-named-entity-recognition-english.

Bảng 4.3: Kết quả từ nhiệm vụ nhận dạng thực thể được đặt tên. Hai số liệu ở đây là điểm F1 và độ chính xác. BERT có dấu sao biểu thị BERT do thư viện Flair cung cấp. Kết quả tốt nhất được in đậm.

Mô hình	Độ chính xác của điểm F1
Flair (cơ sở)	0,9235 0,8578
wac2vec	0,7381 0,5849
GloVe	0,8882 0,7988
BERT	0,6304 0,4603
BERT*	0,9131 0,8401
wac2vec + Flair 0,9017 wac2vec +	0,8209
GloVe 0,8508 wac2vec + BERT	0,7403
0,7719 wac2vec + BERT* 0,9113	0,6284
	0,8371

được sử dụng đúng ở cấp độ câu. Giống như trong cụm từ chunking task, wac2vec by bản thân nó thực hiện kém nhiệm vụ này.

4.4 Sự tương đồng của từ

Độ tương đồng của từ cho biết mức độ giống nhau của hai từ trong không gian vectơ. Điều này hữu ích vì các nhúng tiêu chuẩn và mô hình wac2vec của chúng tôi được thể hiện dưới dạng các vectơ, với giả định rằng hai vectơ gần nhau trong không gian vectơ là tương tự về mặt ngữ nghĩa. Đây là nhiệm vụ đơn giản nhất trong tất cả các nhiệm vụ của chúng tôi, vì nó chỉ liên quan đến tính toán mức độ gần nhau của hai nhúng cho một cặp từ trong không gian vectơ, sử dụng độ tương đồng cosin (giải thích bên dưới).

bờ biển - bờ biển [rất giống nhau]

cây - xe [không giống nhau]

4.4.1 Nhiệm vụ & Thủ tục

Nhiệm vụ này bao gồm dự đoán sự tương đồng về mặt ngữ nghĩa của các cặp từ trong hai tập dữ liệu, WordSim-353 và SimLex-999 (xem Phần 3.1).6

Chúng tôi đã thử nghiệm một số kết hợp nhúng, sử dụng cả GloVe và BERT, để so sánh lợi ích tương đối của việc áp dụng wac2vec cho các mô hình khác nhau. Xem Bảng 4.4 để có danh sách tất cả các nhúng đã được thử nghiệm. Độ tương tự cosin giữa các nhúng của hai từ trong một cặp được tính bằng:

$$\frac{vec1 \cdot vec2}{||vec1|| ||vec2||} \tag{4.7}$$

trong đó vec1 là nhúng của từ đầu tiên trong cặp và vec2 là nhúng của từ thứ hai trong cặp.

4.4.2 Số liệu

Tiếp theo công trình trước đó [36], thước đo của chúng tôi cho nhiệm vụ này là tương quan Spearman, như một cách để tổng hợp các điểm tương đồng cosin giữa các từ trong các tập dữ liệu. Nhiệm vụ này sử dụng phương pháp spearmanr được cung cấp bởi thư viện SciPy Python và tính toán sự tương quan Spearman rs trên các từ được xếp hạng giống nhau rx được cung cấp bởi một nhúng và các điểm tương đồng được xếp hạng được cung cấp bởi tập dữ liệu. Trong theo phương trình sau, cov(rx, ry) là hiệp phương sai của rx và ry, và σ rx và σ ry là độ lệch chuẩn của rx và ry tương ứng.

$$rs = \frac{cov(rx, ry)}{\sigma rx \sigma ry} \tag{4.8}$$

6Chúng tôi đã sử dụng mã từ dự án này để tính toán mức độ tương đồng giữa các cặp từ: https://github.com/recski/wordsim

Bảng 4.4: Kết quả từ nhiệm vụ so sánh từ. Nửa trên của kết quả tương ứng với tập dữ liệu SimLex-999 và nửa dưới hiển thị kết quả trên tập dữ liệu WordSim-353. Chỉ số được sử dụng trong nhiệm vụ này là tương quan Spearman. Kết quả tốt nhất được in đậm.

Người mẫu	Người cầm giáo
wac2vec 0.0198 Găng tay 0.3392 BERT 0,1582 wac2vec + GloVe 0.0585 wac2vec + BERT 0.0618	
wac2vec 0.1565 Găng tay 0.6312 BERT 0.3750 wac2vec + GloVe 0.2503 wac2vec + BERT 0.2990	

4.4.3 Kết quả

Bảng 4.4 cho thấy kết quả cho nhiệm vụ này. Như có thể thấy, đối với cả SimLex-999 tập dữ liệu và tập dữ liệu WordSim-999, thêm wac2vec vào bất kỳ một trong hai phân phối nhúng quốc tế không giúp cải thiện kết quả.

Một điều thú vị trong những kết quả này là BERT bị GloVe vượt trội hơn hẳn, mặc dù nó thiết lập trạng thái nghệ thuật cho nhiều nhiệm vụ NLP khác. Chúng tôi suy đoán điều này là do bản chất của BERT—một nhiệm vụ như thế này hoàn toàn nằm trên từ cấp độ, thay vì cấp độ câu, nơi BERT có thể hoạt động tốt nhất. Không ngờ rằng, wac2vec không thực hiện tốt nhiệm vụ này. Có vẻ như nó không nắm bắt được biểu diễn trong không gian vectơ cũng như chúng ta nghĩ.

4.5 Thảo luận về kết quả

Trong chương này, chúng tôi đã đánh giá wac2vec và sự kết hợp của wac2vec và một tra- nhúng phân phối theo truyền thống, trên một số nhiệm vụ NLP: đối thoại trực quan, cụm từ

phân nhóm, nhận dạng thực thể được đặt tên và độ tương đồng của từ.

Wac2vec, một mô hình nhúng được áp dụng trực tiếp vào các từ, hoạt động tốt trên
của riêng nó trong cuộc đối thoại trực quan, một nhiệm vụ chứa đựng một số lượng lớn các thông tin cụ thể, trực quan
từ ngữ có căn cứ. Hơn nữa, hiệu suất của nhúng từ ngữ truyền thống là
được cải thiện khi kết hợp với nhúng wac2vec. Chúng tôi tin rằng điều này chứng minh
wac2vec, được đào tạo trên dữ liệu trực quan, làm phong phú thêm các nhúng truyền thống,
được đào tạo trên dữ liệu văn bản.

Trong khi wac2vec thực hiện tốt và cải thiện hiệu suất cho phân phối
nhúng vào nhiệm vụ có cơ sở trực quan của chúng tôi, nó không làm như vậy đối với thực thể được đặt tên
nhiệm vụ nhận dạng và phân nhóm cụm từ, thường là các bài kiểm tra cú pháp (được đặt tên
thực thể thường là cụm danh từ không nhất thiết biểu thị những thứ cụ thể).
thực tế là wac2vec không hoạt động tốt khi hoạt động độc lập cũng như không có xu hướng cải thiện hiệu suất
khi được nối với các nhúng truyền thống cho thấy wac2vec không học được
cú pháp. Những kết quả này có thể hiểu được, vì wac2vec đưa ra giả định về từ vựng
tính độc lập–bộ phân loại cho mỗi từ được đào tạo độc lập với tất cả các từ khác.
Một lý do khác cho kết quả kém của hai nhiệm vụ này có thể là các tập dữ liệu mà chúng
sử dụng chứa ngôn ngữ ít căn cứ hơn nhiều so với tập dữ liệu được sử dụng trong hội thoại trực quan
nhiệm vụ.

Một điều cần cân nhắc là liệu kích thước nhúng lớn hơn của wac2vec có đóng góp hay không
để có hiệu suất cao hơn, thay vì sự hiện diện của thông tin trực quan quan trọng
trong các vectơ của chúng tôi. Rốt cuộc, trong nhiệm vụ đối thoại trực quan, wac2vec, với kích thước rất lớn
vector 1300, luôn hoạt động tốt hơn cả fastText và GloVe trên tất cả
số liệu. Tuy nhiên, nó hầu như luôn bị BERT, một vectơ nhỏ hơn, đánh bại
ở kích thước 768 và trong ba nhiệm vụ khác, wac2vec có xu hướng thực hiện thấp nhất,
không có cải tiến nào khi kết hợp với nhúng truyền thống. Bởi vì điều này,

chúng tôi tin rằng hiệu suất của wac2vec trong nhiệm vụ đối thoại trực quan, cả riêng lẻ và khi được nối với một nhúng khác, không phải do tính đa chiều cao hơn của nó (trong so sánh với các nhúng khác).

CHƯƠNG 5

KẾT LUẬN

5.1 Chúng tôi đã làm được những gì cho đến nay?

Trong luận án này, chúng tôi đã đánh giá wac2vec, một mô hình dựa trên se-
mantics bằng cách đào tạo các bộ phân loại từ riêng lẻ trên một bộ sưu tập hình ảnh đã tải xuống
từ các công cụ tìm kiếm hình ảnh, sử dụng vốn từ vựng có nguồn gốc từ Thời đại tiếp thu
tập dữ liệu và các tập dữ liệu được sử dụng trong việc đánh giá mô hình của chúng tôi. Chúng tôi đã trích xuất các hệ số
từ các bộ phân loại này để tạo thành các nhúng từ có căn cứ và kết hợp chúng với
nhúng từ truyền thống mô hình hóa ý nghĩa ngữ nghĩa phân phối. Cuối cùng,
chúng tôi đã tiến hành phân tích thực nghiệm trên mô hình của mình để hiểu rõ hơn về những gì nó học được
và liệu nó có làm phong phú thêm các nhúng chỉ có văn bản hay không, bằng cách sử dụng bốn tác vụ NLP: đối thoại trực quan,
phân cụm cụm từ, nhận dạng thực thể được đặt tên và độ tương đồng của từ.

Đối với hầu hết các nhiệm vụ của chúng tôi, việc nối wac2vec với nhúng phân phối
có xu hướng không cải thiện hiệu suất. Điều này không phải là bất ngờ: Tôi phỏng đoán rằng
điều này có thể là do các nhiệm vụ phân cụm cụm từ và nhận dạng thực thể được đặt tên là
chủ yếu là các bài kiểm tra cú pháp, trong khi mô hình của chúng tôi đưa ra giả định về tính độc lập
của các từ; điều này có nghĩa là nó không xử lý tốt các tác vụ cú pháp.

Tuy nhiên, chúng tôi đã quan sát thấy những cải tiến trong nhiệm vụ đối thoại trực quan khi com-
kết hợp wac2vec với cả nhúng GloVe và BERT, so với những
nhúng một mình. Bộ dữ liệu đối thoại trực quan là một nhiệm vụ dựa trên trực quan và

chứa ngôn ngữ cụ thể đề cập đến và mô tả một loạt hình ảnh. Thực tế
rằng hiệu suất được cải thiện trên nhiệm vụ này khi kết hợp wac2vec và truyền thống
nhúng cho thấy wac2vec cung cấp thành công thông tin ngữ nghĩa có cơ sở
không được nắm bắt bởi những nhúng này.

Tóm lại, chúng tôi thấy rằng việc kết hợp wac2vec, được nối đất, với một phân phối
nhúng ngữ nghĩa bổ sung thông tin ngữ nghĩa quan trọng, vì hiệu suất được cải thiện
khi sử dụng cả hai cùng nhau trong một nhiệm vụ trực quan (thay vì sử dụng riêng từng phương pháp).
Điều này có ý nghĩa đối với bất kỳ nhiệm vụ nào có thể sử dụng những từ ngữ cụ thể (ví dụ: đối thoại trực quan,
trả lời câu hỏi trực quan, giải quyết tham chiếu và tương tác với robot). Như ex-
dự kiến, nó không giúp cải thiện hiệu suất trong các tác vụ thiên về cú pháp hơn, cho thấy
WAC không học được bất cứ điều gì về cú pháp; các vectơ wac2vec phải luôn luôn là
được sử dụng kết hợp với các nhúng phân phối thích hợp như GloVe hoặc BERT.

5.2 Hướng đi trong tương lai

Trong công việc tương lai, chúng tôi sẽ thực hiện việc cân nhắc các từ theo mức độ cụ thể. Từ ngữ
trừu tượng hơn sẽ được chú trọng hơn vào nhúng, và những thứ đó
cụ thể hơn sẽ được cân nhắc nhiều hơn về phía wac2vec. Chúng tôi cũng muốn khám phá
cách wac2vec hoạt động với toàn bộ sức mạnh của mô hình biến áp giống BERT; trong này
luận án, chúng tôi chỉ trích xuất nhúng từ cho BERT bằng cách truyền một từ duy nhất tại
một thời gian thông qua mô hình BERT, thay vì truyền toàn bộ một câu, do
hạn chế của phần cứng của chúng tôi. Đây là một cách tiếp cận hợp lệ cho nhiệm vụ tương tự từ
(vì không có câu, chỉ có các từ riêng lẻ), nhưng nhiệm vụ đối thoại trực quan,
sử dụng câu hỏi và câu trả lời, sẽ có lợi nếu chuyển toàn bộ câu vào BERT.

TÀI LIỆU THAM KHẢO

[1] Tanmay Gupta, Alexander Schwing và Derek Hoiem. Vico: Nhúng từ từ các hiện tượng đồng thời trực quan. Trong Biên bản báo cáo Hội nghị quốc tế IEEE về thị giác máy tính, trang 7425–7434, 2019.

[2] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca và Aitor Soroa. Một nghiên cứu về sự tương đồng và liên quan sử dụng các phương pháp phân phối và dựa trên mạng từ. Trong Biên bản báo cáo của HLT-NAACL, trang 19–27, 2009.

[3] Alan Akbik, Duncan Blythe và Roland Vollgraf. Nhúng chuỗi theo ngữ cảnh để dán nhãn chuỗi. Trong Biên bản báo cáo Hội nghị quốc tế lần thứ 27 về Ngôn ngữ học tính toán, trang 1638–1649, 2018.

[4] Laura Barca, Cristina Burani và Lisa S Arduino. Thời gian đặt tên từ và chuẩn mực tâm lý ngôn ngữ cho danh từ tiếng Ý. Phương pháp nghiên cứu hành vi, công cụ và máy tính, 34(3):424–434, 2002.

[5] Emily M. Bender và Alexander Koller. Leo lên tới nlu: Về ý nghĩa, hình thức và sự hiểu biết trong thời đại dữ liệu. Trong Biên bản báo cáo của Hội nghị thường niên lần thứ 58 của Hiệp hội Ngôn ngữ học tính toán, 2020.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin và Tomas Mikolov. Làm giàu các vectơ từ bằng thông tin từ phụ. Giao dịch của Hiệp hội Ngôn ngữ học tính toán, 5:135–146, 2017.

[7] Elia Bruni, Gemma Boleda, Marco Baroni và Nam-Khanh Tran. Ngữ nghĩa phân phối trong technicolor. Trong Biên bản báo cáo của Hội nghị thường niên lần thứ 50 của Hiệp hội Ngôn ngữ học tính toán: Bài báo dài - Tập 1, trang 136–145. Hiệp hội Ngôn ngữ học tính toán, 2012.

[8] Joyce Y Chai, Rui Fang, Changsong Liu và Lanbo She. Nền tảng ngôn ngữ cộng tác hướng tới đối thoại giữa người và rô-bốt có vị trí. Tạp chí AI, 37(4):32–45, 2016.

[9] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely và Yoav Artzi. Touchdown: Điều hướng ngôn ngữ tự nhiên và lý luận không gian trong môi trường đường phố trực quan. Trong Biên bản báo cáo của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 12538–12547, 2019.

[10] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh và Dhruv Batra. Đối thoại trực quan. Trong Biên bản báo cáo Hội nghị IEEE về Tầm nhìn máy tính và Nhận dạng mẫu, trang 326–335, 2017.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee và Kristina Toutanova. Bert: Đào tạo trước các bộ biến đổi song hướng sâu để hiểu ngôn ngữ. Bản in trước arXiv arXiv:1810.04805, 2018.

[12] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman và Eytan Ruppin. Đặt tìm kiếm trong ngữ cảnh: Khái niệm được xem xét lại. ACM Transactions on information systems, 20(1):116–131, 2002.

[13] John R Firth. Tóm tắt lý thuyết ngôn ngữ học, 1930-1955. Nghiên cứu về ngôn ngữ học phân tích, 1957.

[14] Aurelie Herbelot. Những gì có trong một văn bản, những gì không có, và điều này liên quan gì đến ngữ nghĩa từ vựng. Trong Biên bản Hội nghị quốc tế lần thứ 10 về ngữ nghĩa tính toán (IWCS 2013)–Bài báo ngắn, trang 321–327, 2013.

[15] Felix Hill, Roi Reichart và Anna Korhonen. Simlex-999: Đánh giá các mô hình ngữ nghĩa với ước tính độ tương đồng (thực sự). Ngôn ngữ học tính toán, 41(4):665–695, 2015.

[16] Elizabeth Jefferies, Karalyn Patterson, Roy W Jones và Matthew A Lam-bon Ralph. Hiểu các từ cụ thể và trừu tượng trong chứng mất trí ngữ nghĩa.
Tâm lý học thần kinh, 23(4):492, 2009.

[17] Casey Kennington và David Schlangen. Học tập đơn giản và ứng dụng sáng tác các ý nghĩa từ ngữ có cơ sở nhận thức để giải quyết tham chiếu gia tăng. Trong Biên bản báo cáo của Hội nghị thường niên lần thứ 53 của Hiệp hội Ngôn ngữ học tính toán và Hội nghị chung quốc tế lần thứ 7 về Xử lý ngôn ngữ tự nhiên (Tập 1: Bài báo dài), trang 292–301, 2015.

[18] Douwe Kiela và L´eon Bottou. Học nhúng hình ảnh bằng mạng nơ-ron tích chập để cải thiện ngữ nghĩa đa phương thức. Trong Biên bản báo cáo Hội nghị năm 2014 về Phương pháp thực nghiệm trong Xử lý ngôn ngữ tự nhiên (EMNLP), trang 36–45, 2014.

[19] Jamie Kiros, William Chan và Geoffrey Hinton. Hiểu ngôn ngữ minh họa: Nền tảng trực quan quy mô lớn với tìm kiếm hình ảnh. Trong Biên bản báo cáo của Hội nghị thường niên lần thứ 56 của Hiệp hội Ngôn ngữ học tính toán (Tập 1: Bài báo dài), trang 922–933, 2018.

[20] Ryan Kiros, Ruslan Salakhutdinov và Richard S Zemel. Thống nhất nhúng ngữ nghĩa thị giác với các mô hình ngôn ngữ thần kinh đa phương thức. Bản in trước arXiv arXiv:1411.2539, 2014.

[21] Judith F Kroll và Jill S Merves. Truy cập từ vựng cho các từ cụ thể và trừu tượng. Tạp chí Tâm lý học Thực nghiệm: Học tập, Trí nhớ và Nhận thức, 12(1):92, 1986.

[22] Victor Kuperman, Hans Stadthagen-Gonzalez và Marc Brysbaert. Xếp hạng độ tuổi tiếp thu cho 30.000 từ tiếng Anh. Phương pháp nghiên cứu hành vi, 44(4):978–990, 2012.

[23] Angeliki Lazaridou, Elia Bruni và Marco Baroni. Đây có phải là wampimuk không? ánh xạ liên phương thức giữa ngữ nghĩa phân phối và thế giới thị giác. Trong Biên bản báo cáo của Hội nghị thường niên lần thứ 52 của Hiệp hội Ngôn ngữ học tính toán (Tập 1: Bài báo dài), trang 1403–1414, 2014.

[24] Alessandro Lenci. Các mô hình phân phối ý nghĩa của từ. Đánh giá hàng năm của Ngôn ngữ học, 4:151–171, 2018.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar và C Lawrence Zitnick. Microsoft coco: Các đối tượng chung trong ngữ cảnh. Trong hội nghị châu Âu về tầm nhìn máy tính, trang 740–755. Springer, 2014.

[26] Max M Louwerse. Sự phụ thuộc lẫn nhau của biểu tượng trong nhận thức biểu tượng và nhận thức hiện thân. Chủ đề trong Khoa học Nhận thức, 3(2):273–302, 2011.

[27] David G Lowe. Nhận dạng đối tượng từ các đặc điểm bất biến theo tỷ lệ cục bộ. Trong Biên bản báo cáo của hội nghị quốc tế IEEE lần thứ bảy về thị giác máy tính, tập 2, trang 1150–1157. Ieee, 1999.

[28] Jiasen Lu, Dhruv Batra, Devi Parikh và Stefan Lee. Vilbert: Đào tạo trước các biểu diễn ngôn ngữ thị giác không phụ thuộc vào nhiệm vụ cho các nhiệm vụ ngôn ngữ và thị giác. Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, trang 13–23, 2019.

[29] Andy L¨ucking, Robin Cooper, Staffan Larsson và Jonathan Ginzburg. Phân phối không đủ: tiến xa hơn. Trong Biên bản Hội thảo lần thứ sáu về Ngôn ngữ tự nhiên và Khoa học máy tính, trang 1–10, 2019.

[30] Laurens van der Maaten và Geoffrey Hinton. Trực quan hóa dữ liệu bằng t-sne. Tạp chí nghiên cứu máy học, 9 (tháng 11): 2579–2605, 2008.

[31] Daniela Massiceti, Puneet K Dokania, N Siddharth và Philip HS Torr. Đối thoại trực quan không có thị giác hoặc đối thoại. Bản in trước arXiv arXiv:1812.06417, 2018.

[32] Tomas Mikolov, Kai Chen, Greg Corrado và Jeffrey Dean. Ước tính hiệu quả các biểu diễn từ trong không gian vectơ. Bản in trước arXiv arXiv:1301.3781, 2013.

[33] Daniele Moro, Stacy Black và Casey Kennington. Soạn thảo và nhúng mô hình từ-như-phân-loại của ngữ nghĩa cơ bản. Bản in trước arXiv arXiv:1911.03283, 2019.

[34] Jeffrey Pennington, Richard Socher và Christopher Manning. Glove: Các vectơ toàn cục để biểu diễn từ. Trong Biên bản báo cáo hội nghị năm 2014 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên (EMNLP), trang 1532–1543, 2014.

[35] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee và Luke Zettlemoyer. Biểu diễn từ ngữ theo ngữ cảnh sâu. Bản in trước arXiv arXiv:1802.05365, 2018.

[36] G'abor Andr'as Recski, Eszter Ikl'odi, Katalin Anna Pajkossy, và Andras Kornai. Đo lường sự tương đồng về mặt ngữ nghĩa của các từ bằng cách sử dụng mạng lưới khái niệm. Hiệp hội Ngôn ngữ học tính toán, 2016.

[37] Erik F Sang và Sabine Buchholz. Giới thiệu về nhiệm vụ chung conll-2000: Phân mảnh. Bản in trước của arXiv cs/0009008, 2000.

[38] Erik F Sang và Fien De Meulder. Giới thiệu về nhiệm vụ chung conll-2003: Nhận dạng thực thể có tên độc lập với ngôn ngữ. Bản in trước arXiv cs/0306050, 2003.

[39] David Schlangen, Sina Zarrieß và Casey Kennington. Giải quyết các tham chiếu đến các đối tượng trong ảnh bằng mô hình từ ngữ làm bộ phân loại. Bản in trước arXiv arXiv:1510.02125, 2015.

[40] Erich Schubert, J¨org Sander, Martin Ester, Hans Peter Kriegel và Xiaowei Xu. Xem lại Dbscan, xem lại: tại sao và làm thế nào bạn (vẫn) nên sử dụng dbscan. ACM Transactions on Database Systems (TODS), 42(3):19, 2017.

[41] Filip Smol´ık. Khả năng hình dung danh từ tạo điều kiện thuận lợi cho việc tiếp thu số nhiều: phân tích sự sống còn của sự xuất hiện số nhiều ở trẻ em. Tạp chí nghiên cứu tâm lý ngôn ngữ, 43(4):335–350, 2014.

[42] Mingxing Tan và Quoc V Le. Efficientnet: Xem xét lại việc mở rộng mô hình cho mạng nơ-ron tích chập. Bản in trước arXiv arXiv:1905.11946, 2019.

[43] Serge Thill và Katherine E Twomey. Những gì bên trong mới là quan trọng: Một tường thuật có cơ sở về quá trình tiếp thu và phát triển khái niệm. Biên giới trong tâm lý học, 7:402, 2016.

[44] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone và Raymond J Mooney. Học ngữ nghĩa ngôn ngữ học đa phương thức có cơ sở bằng cách chơi “tôi là gián điệp”. Trong IJCAI, trang 3477–3483, 2016.

[45] Peter D Turney và Patrick Pantel. Từ tần suất đến ý nghĩa: Mô hình không gian vectơ của ngữ nghĩa. Tạp chí nghiên cứu trí tuệ nhân tạo, 37:141–188, 2010.

[46] Lichen Yu, Patrick Poirson, Shan Yang, Alexander C Berg, và Tamara L Berg. Mô hình hóa ngữ cảnh trong biểu thức tham chiếu. Trong Hội nghị Châu Âu về Thị giác máy tính, trang 69–85. Springer, 2016.

[47] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba và Sanja Fidler. Căn chỉnh sách và phim: Hướng tới các giải thích trực quan giống như câu chuyện bằng cách xem phim và đọc sách. Trong Biên bản báo cáo hội nghị quốc tế IEEE về thị giác máy tính, trang 19–27, 2015.