Cải thiện khả năng hiểu ngôn ngữ bằng cách đào tạo trư ớc sáng tạo

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

trừu tư ợng

Hiểu ngôn ngữ tự nhiên bao gồm một loạt các nhiệm vụ đa dạng như trình bày theo văn bản, trả lời câu hỏi, đánh giá sự tư ơ ng đồng về ngữ nghĩa và phân loại tài liệu. Mặc dù có rất nhiều kho văn bản lớn không đư ợc gắn nhãn, như ng dữ liệu đư ợc gắn nhãn để học các nhiệm vụ cụ thể này lại khan hiếm, khiến các mô hình đư ợc đào tạo phân biệt gặp khó khăn để thực hiện đầy đủ. Chúng tôi chứng minh rằng những lợi ích lớn đối với các nhiệm vụ này có thể đư ợc thực hiện bằng cách đào tạo trư ớc một cách tổng quát một mô hình ngôn ngữ trên một kho văn bản không đư ợc gắn nhãn đa dạng, sau đó là tinh chỉnh phân biệt đối xử trên từng nhiệm vụ cụ thể. Trái ngư ợc với các phư ơ ng pháp trư ớc đây, chúng tôi sử dụng các phép biến đổi đầu vào nhận biết tác vụ trong quá trình tinh chỉnh để đạt đư ợc sự chuyển giao hiệu quả trong khi yêu cầu những thay đổi tối thiểu đối với kiến trúc mô hình. Chúng tôi chứng minh tính hiệu quả của phư ơ ng pháp tiếp cận của chúng tôi trên nhiều tiêu chuẩn đánh giá mức độ hiểu ngôn ngữ tự nhiên. Mô hình bất khả tri về nhiệm vụ chung của chúng tôi vư ợt trội so với các mô hình đư ợc đào tạo chuyên biệt sử dụng các kiến trúc đư ợc tạo riêng cho từng nhiệm vụ, cải thiện đáng kể dựa trên trạng thái hiện đại ở 9 trong số 12 nhiệm vụ đư ợc nghiên cứu. Ví dụ: chúng tôi đạt đư ợc mức cải thiện tuyệt đối là 8,9% đối với khả năng lập luận thông thư ờng (Thử nghiệm Cloze Câu chuyện), 5,7% đối với trả lời câu hỏi (RACE) và 1,5% đối với yêu cầu văn bản (MultiNLI).

1. Giới thiệu

Khả năng học hiệu quả từ văn bản thô là rất quan trọng để giảm bớt sự phụ thuộc vào việc học có giám sát trong xử lý ngôn ngữ tự nhiên (NLP). Hầu hết các phư ơ ng pháp học sâu đều yêu cầu một lư ợng lớn dữ liệu đư ợc gắn nhãn thủ công, điều này hạn chế khả năng ứng dụng của chúng trong nhiều lĩnh vực thiếu tài nguyên chú thích [61]. Trong những tình huống này, các mô hình có thể tận dụng thông tin ngôn ngữ từ dữ liệu chư a đư ợc gắn nhãn cung cấp một giải pháp thay thế có giá trị để thu thập thêm chú thích, có thể tốn thời gian và tốn kém. Hơ n nữa, ngay cả trong những trư ờng hợp có sẵn sự giám sát đáng kể, việc học các biểu diễn tốt theo kiểu không giám sát có thể giúp tăng hiệu suất đáng kể. Bằng chứng thuyết phục nhất cho điều này cho đến nay là việc sử dụng rộng rãi các nhúng từ đư ợc đào tạo trư ớc [10, 39, 42] để cải thiện hiệu suất trên một loạt các nhiệm vụ NLP [8, 11, 26, 45].

Tuy nhiên, việc tận dụng nhiều hơ n thông tin ở cấp độ từ từ văn bản không đư ợc gắn nhãn là một thách thức vì hai lý do chính. Đầu tiên, không rõ loại mục tiêu tối ư u hóa nào hiệu quả nhất trong việc học các biểu diễn văn bản hữu ích cho việc chuyển giao. Nghiên cứu gần đây đã xem xét các mục tiêu khác nhau như mô hình hóa ngôn ngữ [44], dịch máy [38] và sự gắn kết diễn ngôn [22], với mỗi phư ơ ng pháp vư ợt trội hơ n các phư ơ ng pháp khác trong các nhiệm vụ khác nhau.1 Thứ hai, không có sự đồng thuận về cách hiệu quả nhất để chuyển các biểu diễn đã học này sang nhiệm vụ mục tiêu. Các kỹ thuật hiện có liên quan đến sự kết hợp của việc thực hiện các thay đổi theo nhiệm vụ cụ thể đối với kiến trúc mô hình [43, 44], sử dụng các sơ đồ học tập phức tạp [21] và thêm các mục tiêu học tập phụ trợ [50]. Những điều không chắc chắn này đã gây khó khăn cho việc phát triển các phư ơ ng pháp học bán giám sát hiệu quả để xử lý ngôn ngữ.

[1] https://gluebenchmark.com/leaderboard

Trong bài báo này, chúng tôi khám phá một cách tiếp cận bán giám sát cho các nhiệm vụ hiểu ngôn ngữ bằng cách sử dụng kết hợp đào tạo trư ớc không giám sát và tinh chỉnh có giám sát. Mục tiêu của chúng tôi là tìm hiểu một biểu diễn phổ quát chuyển với ít sự thích ứng với nhiều nhiệm vụ. Chúng tôi giả định quyền truy cập vào một kho văn bản lớn chư a đư ợc gắn nhãn và một số bộ dữ liệu với các ví dụ đào tạo đư ợc chú thích thủ công (tác vụ mục tiêu). Quá trình thiết lập của chúng tôi không yêu cầu các tác vụ mục tiêu này phải ở trong cùng một miền với kho văn bản chư a đư ợc gắn nhãn. Chúng tôi sử dụng quy trình đào tạo hai giai đoạn. Đầu tiên, chúng tôi sử dụng mục tiêu mô hình hóa ngôn ngữ trên dữ liệu chư a đư ợc gắn nhãn để tìm hiểu các tham số ban đầu của mô hình mạng thần kinh. Sau đó, chúng tôi điều chỉnh các tham số này cho một nhiệm vụ mục tiêu bằng cách sử dụng mục tiêu đư ợc giám sát tư ơ ng ứng.

Đối với kiến trúc mô hình của chúng tôi, chúng tôi sử dụng Transformer [62], đã đư ợc chứng minh là thực hiện tốt các tác vụ khác nhau như dịch máy [62], tạo tài liệu [34] và phân tích cú pháp [29].
Lựa chọn mô hình này cung cấp cho chúng tôi bộ nhớ có cấu trúc hơ n để xử lý các phụ thuộc dài hạn trong văn bản, so với các lựa chọn thay thế như mạng lặp lại, dẫn đến hiệu suất truyền tải mạnh mẽ qua nhiều tác vụ khác nhau. Trong quá trình chuyển, chúng tôi sử dụng các điều chỉnh đầu vào dành riêng cho nhiệm vụ bắt nguồn từ các phư ơ ng pháp tiếp cận kiểu truyền tải [52], xử lý đầu vào văn bản có cấu trúc dư ới dạng một chuỗi mã thông báo liền kề duy nhất. Như chúng tôi chứng minh trong các thử nghiệm của mình, những điều chỉnh này cho phép chúng tôi tinh chỉnh một cách hiệu quả với những thay đổi tối thiểu đối với kiến trúc của mô hình đư ợc đào tạo trư ớc.

Chúng tôi đánh giá cách tiếp cận của mình trên bốn loại nhiệm vụ hiểu ngôn ngữ – suy luận ngôn ngữ tự nhiên, trả lời câu hỏi, tư ơ ng đồng ngữ nghĩa và phân loại văn bản. Mô hình bất khả tri về nhiệm vụ chung của chúng tôi hoạt động tốt hơ n các mô hình đư ợc đào tạo chuyên biệt sử dụng các kiến trúc đư ợc tạo riêng cho từng nhiệm vụ, cải thiện đáng kể dựa trên trạng thái hiện đại ở 9 trong số 12 nhiệm vụ đư ợc nghiên cứu. Chẳng hạn, chúng tôi đạt đư ợc mức cải thiện tuyệt đối là 8,9% đối với khả năng lập luận thông thư ờng (Thử nghiệm Cloze Câu chuyện) [40], 5,7% đối với trả lời câu hỏi (RACE) [30], 1,5% đối với yêu cầu văn bản (MultiNLI) [66] và 5,5% đối với gần đây đã giới thiệu điểm chuẩn đa tác vụ GLUE [64]. Chúng tôi cũng đã phân tích các hành vi zero-shot của mô hình đư ợc đào tạo trư ớc trên bốn cài đặt khác nhau và chứng minh rằng nó thu đư ợc kiến thức ngôn ngữ hữu ích cho các tác vụ tiếp theo.

2 công việc liên quan

Học bán giám sát cho NLP Công việc của chúng tôi nói chung thuộc thể loại học bán giám sát cho ngôn ngữ tự nhiên. Mô hình này đã thu hút đư ợc sự quan tâm đáng kể, với các ứng dụng cho các tác vụ như ghi nhãn trình tự [24, 33, 57] hoặc phân loại văn bản [41, 70]. Các phư ơ ng pháp tiếp cận sớm nhất đã sử dụng dữ liệu chư a đư ợc gắn nhãn để tính toán thống kê cấp độ từ hoặc cấp độ cụm từ, sau đó đư ợc sử dụng làm tính năng trong mô hình đư ợc giám sát [33]. Trong vài năm qua, các nhà nghiên cứu đã chứng minh những lợi ích của việc sử dụng nhúng từ [11, 39, 42], đư ợc đào tạo trên kho ngữ liệu không đư ợc gắn nhãn, để cải thiện hiệu suất trong nhiều nhiệm vụ khác nhau [8, 11, 26, 45]. Tuy nhiên, những cách tiếp cận này chủ yếu chuyển thông tin cấp độ từ, trong khi chúng tôi nhắm đến việc nắm bắt ngữ nghĩa cấp cao hơ n.

Các cách tiếp cận gần đây đã điều tra việc học và sử dụng nhiều hơ n ngữ nghĩa cấp độ từ từ dữ liệu chư a đư ợc gắn nhãn. Các nhúng ở cấp độ cụm từ hoặc cấp độ câu, có thể đư ợc đào tạo bằng cách sử dụng kho văn bản không đư ợc gắn nhãn, đã đư ợc sử dụng để mã hóa văn bản thành các biểu diễn véc-tơ phù hợp cho các nhiệm vụ mục tiêu khác nhau [28, 32, 1, 36, 22, 12, 56, 31].

Đào tạo trư ớc không giám sát Đào tạo trư ớc không giám sát là một trư ờng hợp đặc biệt của học bán giám sát trong đó mục tiêu là tìm một điểm khởi tạo tốt thay vì sửa đổi mục tiêu học có giám sát. Các công trình ban đầu đã khám phá việc sử dụng kỹ thuật này trong phân loại hình ảnh [20, 49, 63] và các tác vụ hồi quy [3]. Nghiên cứu tiếp theo [15] đã chứng minh rằng đào tạo trư ớc hoạt động như một sơ đồ chính quy hóa, cho phép khái quát hóa tốt hơ n trong các mạng lư ới thần kinh sâu. Trong công việc gần đây, phư ơ ng pháp này đã đư ợc sử dụng để giúp đào tạo mạng lư ới thần kinh sâu về các tác vụ khác nhau như phân loại hình ảnh [69], nhận dạng giọng nói [68], phân định thực thể [17] và dịch máy [48].

Dòng công việc gần nhất với chúng tôi liên quan đến việc đào tạo trư ớc một mạng lư ới thần kinh bằng cách sử dụng mục tiêu mô hình hóa ngôn ngữ và sau đó tinh chỉnh nó theo nhiệm vụ mục tiêu với sự giám sát. Đại và cộng sự. [13] và Howard và Ruder [21] theo phư ơ ng pháp này để cải thiện việc phân loại văn bản. Tuy nhiên, mặc dù giai đoạn tiền đào tạo giúp nắm bắt một số thông tin ngôn ngữ, việc sử dụng các mô hình LSTM của họ hạn chế khả năng dự đoán của họ trong phạm vi ngắn. Ngư ợc lại, sự lựa chọn của chúng tôi về các mạng máy biến áp cho phép chúng tôi nắm bắt đư ợc cấu trúc ngôn ngữ phạm vi dài hơ n, như đã đư ợc chứng minh trong các thí nghiệm của chúng tôi. Hơ n nữa, chúng tôi cũng chứng minh tính hiệu quả của mô hình của chúng tôi đối với nhiều nhiệm vụ hơ n bao gồm suy luận ngôn ngữ tự nhiên, phát hiện diễn giải và hoàn thành câu chuyện. Các cách tiếp cận khác [43, 44, 38] sử dụng các biểu diễn ẩn từ một

mô hình dịch máy hoặc ngôn ngữ đư ợc đào tạo trư ớc làm các tính năng phụ trợ trong khi đào tạo mô hình đư ợc giám sát về tác vụ đích. Điều này liên quan đến một số lư ợng đáng kể các tham số mới cho từng tác vụ mục tiêu riêng biệt, trong khi chúng tôi yêu cầu những thay đổi tối thiểu đối với kiến trúc mô hình của mình trong quá trình chuyển giao.

Các mục tiêu đào tạo phụ trợ Thêm các mục tiêu đào tạo không giám sát phụ trợ là một hình thức thay thế của học bán giám sát. Công việc ban đầu của Collobert và Weston [10] đã sử dụng nhiều tác vụ NLP phụ trợ như gắn thẻ POS, phân đoạn, nhận dạng thực thể đư ợc đặt tên và mô hình hóa ngôn ngữ để cải thiện việc ghi nhãn vai trò ngữ nghĩa. Gần đây hơ n, Rei [50] đã thêm mục tiêu mô hình hóa ngôn ngữ phụ trợ vào mục tiêu nhiệm vụ mục tiêu của họ và chứng minh hiệu suất đạt đư ợc đối với các nhiệm vụ ghi nhãn trình tự. Các thử nghiệm của chúng tôi cũng sử dụng một mục tiêu phụ trợ, như ng như chúng tôi chỉ ra, quá trình đào tạo trư ớc không đư ợc giám sát đã học đư ợc một số khía cạnh ngôn ngữ liên quan đến các nhiệm vụ mục tiêu.

3 Khung

Quy trình đào tạo của chúng tôi bao gồm hai giai đoạn. Giai đoạn đầu là học mô hình ngôn ngữ dung lư ợng cao trên kho ngữ liệu lớn. Tiếp theo là giai đoạn tinh chỉnh, trong đó chúng tôi điều chỉnh mô hình cho một nhiệm vụ phân biệt đối xử với dữ liệu đư ợc dán nhãn.

3.1 Đào tạo trư ớc không giám sát

Đư a ra một kho mã thông báo không đư ợc giám sát U = {u1, . . . , un}, chúng tôi sử dụng mục tiêu mô hình hóa ngôn ngữ tiêu chuẩn để tối đa hóa khả năng sau:

$$L1(U) = \sum_{i} \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \theta) \quad (1)$$

trong đó k là kích thư ớc của cửa sổ ngữ cảnh và xác suất có điều kiện P đư ợc mô hình hóa bằng mạng thần kinh có tham số θ. Các tham số này đư ợc đào tạo bằng cách sử dụng giảm dần độ dốc ngẫu nhiên [51].

Trong các thử nghiệm của mình, chúng tôi sử dụng bộ giải mã Transformer nhiều lớp [34] cho mô hình ngôn ngữ, đây là một biến thể của biến áp [62]. Mô hình này áp dụng thao tác tự chú ý nhiều đầu đối với mã thông báo ngữ cảnh đầu vào, theo sau là các lớp chuyển tiếp nguồn cấp dữ liệu theo vị trí để tạo phân phối đầu ra trên mã thông báo mục tiêu:

$$h_0 = UW_e + W_p$$
$$h_l = \text{transformer\_block}(h_{l-1}) \quad \forall i \in [1, n]$$
$$P(u) = \text{softmax}(h_n W_e^T) \quad (2)$$

trong đó U = (u −k, . . . , u −1) là vectơ ngữ cảnh của mã thông báo, n là số lớp, We là ma trận nhúng mã thông báo và Wp là ma trận nhúng vị trí.

3.2 Tinh chỉnh có giám sát

Sau khi đào tạo mô hình với mục tiêu trong biểu thức. 1, chúng tôi điều chỉnh các tham số cho tác vụ mục tiêu đư ợc giám sát. Chúng tôi giả sử tập dữ liệu có nhãn C, trong đó mỗi phiên bản bao gồm một chuỗi mã thông báo đầu vào, $x^1, \ldots, x^m$, cùng với nhãn y. Các đầu vào đư ợc chuyển qua mô hình đư ợc đào tạo trư ớc của chúng tôi để thu kích hoạt của khối máy biến áp cuối cùng h_l^m, đư ợc sau đó đư ợc đư a vào lớp đầu ra tuyến tính đư ợc thêm vào với tham số Wy để dự đoán y:

$$P(y | x^1, \ldots, x^m) = \text{softmax}(h_l^m W_y). \quad (3)$$

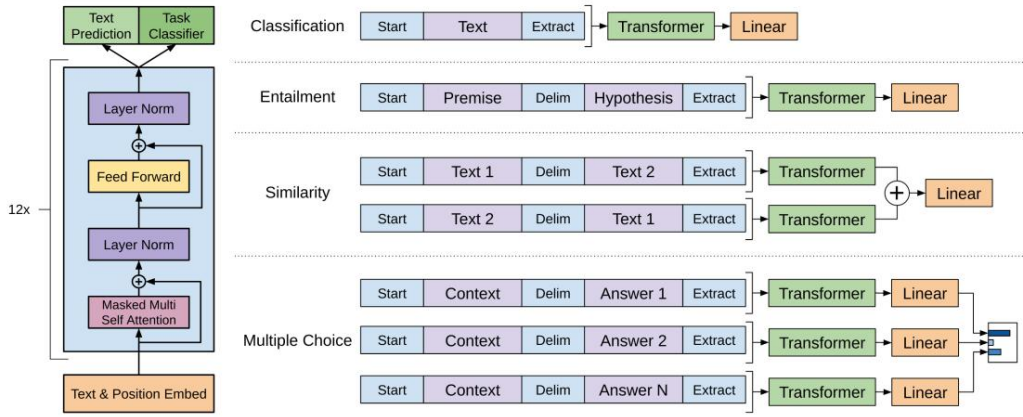
Điều này mang lại cho chúng tôi mục tiêu sau để tối đa hóa:

$$L2(C) = \sum_{(x,y)} \log P(y | x^1, \ldots, x^m). \quad (4)$$

Chúng tôi cũng nhận thấy rằng việc bao gồm mô hình hóa ngôn ngữ như một mục tiêu phụ trợ để tinh chỉnh đã giúp học tập bằng cách (a) cải thiện khả năng khái quát hóa của mô hình đư ợc giám sát và (b) tăng tốc độ hội tụ. Điều này phù hợp với công việc trư ớc đó [50, 43], ngư ời cũng đã quan sát thấy hiệu suất đư ợc cải thiện với mục tiêu phụ trợ như vậy. Cụ thể, chúng tôi tối ư u hóa mục tiêu sau (với trọng số λ):

$$L3(C) = L2(C) + \lambda \times L1(C) \quad (5)$$

Nhìn chung, các tham số bổ sung duy nhất mà chúng tôi yêu cầu trong quá trình tinh chỉnh là Wy và các phần nhúng cho mã thông báo dấu phân cách (đư ợc mô tả bên dư ới trong Phần 3.3).



Hình 1: (trái) Kiến trúc máy biến áp và mục tiêu đào tạo đư ợc sử dụng trong công việc này. (phải) Chuyển đổi đầu vào để tinh chỉnh các tác vụ khác nhau. Chúng tôi chuyển đổi tất cả các đầu vào có cấu trúc thành chuỗi mã thông báo để đư ợc xử lý bởi mô hình đư ợc đào tạo trư ớc của chúng tôi, tiếp theo là lớp tuyến tính + softmax.

3.3 Chuyển đổi đầu vào theo nhiệm vụ cụ thể

Đối với một số tác vụ, chẳng hạn như phân loại văn bản, chúng ta có thể tinh chỉnh trực tiếp mô hình của mình như mô tả ở trên. Một số nhiệm vụ khác, chẳng hạn như trả lời câu hỏi hoặc trình bày theo văn bản, có các đầu vào có cấu trúc như các cặp câu đư ợc sắp xếp theo thứ tự hoặc bộ ba tài liệu, câu hỏi và câu trả lời. Vì mô hình đư ợc đào tạo trư ớc của chúng tôi đã đư ợc đào tạo về các chuỗi văn bản liền kề, nên chúng tôi yêu cầu một số sửa đổi để áp dụng nó cho các tác vụ này.
Công việc trư ớc đây đã đề xuất các kiến trúc cụ thể cho nhiệm vụ học tập trên các biểu diễn đư ợc chuyển giao [44]. Cách tiếp cận như vậy giới thiệu lại một số lư ợng đáng kể tùy chỉnh theo nhiệm vụ cụ thể và không sử dụng học chuyển giao cho các thành phần kiến trúc bổ sung này. Thay vào đó, chúng tôi sử dụng cách tiếp cận kiểu truyền tải [52], trong đó chúng tôi chuyển đổi các đầu vào có cấu trúc thành một chuỗi có thứ tự mà mô hình đư ợc đào tạo trư ớc của chúng tôi có thể xử lý. Những chuyển đổi đầu vào này cho phép chúng tôi tránh thực hiện những thay đổi lớn đối với kiến trúc trong các tác vụ. Chúng tôi cung cấp một mô tả ngắn gọn về các chuyển đổi đầu vào này bên dư ới và Hình 1 cung cấp một minh họa trực quan. Tất cả các chuyển đổi bao gồm thêm mã thông báo bắt đầu và kết thúc đư ợc khởi tạo ngẫu nhiên (s, e).

Văn bản đòi hỏi Đối với các nhiệm vụ đòi hỏi, chúng tôi nối các chuỗi mã thông báo tiền đề p và giả thuyết h , với một mã thông báo phân cách ($) ở giữa.

Tính tư ơ ng tự Đối với các nhiệm vụ tư ơ ng tự, không có thứ tự vốn có của hai câu đư ợc so sánh.
Để phản ánh điều này, chúng tôi sửa đổi trình tự đầu vào để chứa cả hai thứ tự câu có thể có (có dấu phân cách ở giữa) và xử lý từng thứ một cách độc lập để tạo ra hai biểu diễn trình tự h m đư ợc thêm phần tử khôn ngoan trư ớc khi đư ợc đư a vào lớp đầu ra tuyến tính.

Trả lời câu hỏi và lập luận thông thư ờng Đối với những nhiệm vụ này, chúng tôi đư ợc cung cấp một tài liệu ngữ cảnh z, một câu hỏi q và một tập hợp các câu trả lời có thể có {ak}. Chúng tôi nối bối cảnh tài liệu và câu hỏi với từng câu trả lời có thể có, thêm mã thông báo dấu phân cách ở giữa để nhận [z; q; $; ak]. Mỗi trình tự này đư ợc xử lý độc lập với mô hình của chúng tôi và sau đó đư ợc chuẩn hóa thông qua một lớp softmax để tạo ra phân phối đầu ra cho các câu trả lời có thể.

4 thí nghiệm

4.1 Cài đặt

Đào tạo trư ớc không giám sát Chúng tôi sử dụng bộ dữ liệu BooksCorpus [71] để đào tạo mô hình ngôn ngữ.
Nó chứa hơ n 7.000 cuốn sách độc đáo chư a đư ợc xuất bản thuộc nhiều thể loại bao gồm Phiêu lư u, Giả tư ởng và Lãng mạn. Điều quan trọng, nó chứa các đoạn văn bản liền kề dài, cho phép mô hình tổng quát học cách điều chỉnh thông tin tầm xa. Một bộ dữ liệu thay thế, 1B Word Benchmark, đư ợc sử dụng theo cách tiếp cận tư ơ ng tự, ELMo [44], có cùng kích thư ớc

Bảng 1: Danh sách các tác vụ và bộ dữ liệu khác nhau được sử dụng trong các thử nghiệm của chúng tôi.

Nhiệm vụ	Bộ dữ liệu
Suy luận ngôn ngữ tự nhiên	SNLI [5], MultiNLI [66], Câu hỏi NLI [64], RTE [4], SciTail [25]
Trả lời câu hỏi	RACE [30], Story Cloze [40]
câu giống nhau	MSR Paraphrase Corpus [14], Cặp câu hỏi Quora [9], STS Benchmark [6]
phân loại	Stanford Sentiment Treebank-2 [54], CoLA [65]

như ng đư ợc xáo trộn ở cấp độ câu - phá hủy cấu trúc tầm xa. Mô hình ngôn ngữ của chúng tôi đạt đư ợc độ phức tạp ở mức mã thông báo rất thấp là 18,4 trên kho văn bản này.

Thông số kỹ thuật của mô hình Mô hình của chúng tôi phần lớn tuân theo công việc ban đầu của máy biến áp [62]. Chúng tôi đã đào tạo một máy biến áp chỉ dành cho bộ giải mã 12 lớp với các đầu tự chú ý đư ợc đeo mặt nạ (768 trạng thái chiều và 12 đầu chú ý). Đối với các mạng chuyển tiếp nguồn cấp dữ liệu theo vị trí, chúng tôi đã sử dụng trạng thái bên trong 3072 chiều. Chúng tôi đã sử dụng sơ đồ tối ư u hóa Adam [27] với tốc độ học tập tối đa là 2,5e-4. Tỷ lệ học tập đư ợc tăng tuyến tính từ 0 trong 2000 lần cập nhật đầu tiên và giảm dần về 0 bằng cách sử dụng lịch trình cosine. Chúng tôi đào tạo trong 100 kỷ nguyên trên các lô nhỏ gồm 64 chuỗi 512 mã thông báo liền kề đư ợc lấy mẫu ngẫu nhiên. Do layernorm [2] đư ợc sử dụng rộng rãi trong toàn bộ mô hình nên việc khởi tạo trọng số đơ n giản của N(0, 0,02) là đủ. Chúng tôi đã sử dụng từ vựng mã hóa cặp đôi (BPE) với 40.000 lần hợp nhất [53] và phần còn lại, phần nhúng và phần bỏ qua sự chú ý với tỷ lệ 0,1 để chuẩn hóa. Chúng tôi cũng đã sử dụng một phiên bản sửa đổi của chuẩn hóa L2 đư ợc đề xuất trong [37], với w = 0,01 trên tất cả các trọng số không sai lệch hoặc tăng trọng. Đối với chức năng kích hoạt, chúng tôi đã sử dụng Đơ n vị tuyến tính lỗi Gaussian (GELU) [18]. Chúng tôi đã sử dụng các nhúng vị trí đã học thay vì phiên bản hình sin đư ợc đề xuất trong tác phẩm gốc.
Chúng tôi sử dụng thư viện ftfy2 để làm sạch văn bản thô trong BooksCorpus, chuẩn hóa một số dấu chấm câu và khoảng trắng, đồng thời sử dụng mã thông báo spaCy.3

Tinh chỉnh chi tiết Trừ khi đư ợc chỉ định, chúng tôi sử dụng lại các cài đặt siêu tham số từ đào tạo trư ớc không giám sát. Chúng tôi thêm học sinh bỏ học vào bộ phân loại với tỷ lệ 0,1. Đối với hầu hết các nhiệm vụ, chúng tôi sử dụng tỷ lệ học tập là 6,25e-5 và kích thư ớc lô là 32. Mô hình của chúng tôi hoàn thiện nhanh chóng và 3 kỷ nguyên đào tạo là đủ cho hầu hết các trư ờng hợp. Chúng tôi sử dụng lịch trình phân rã tỷ lệ học tập tuyến tính với khởi động trên 0,2% đào tạo. λ đư ợc đặt thành 0,5.

4.2 Tinh chỉnh có giám sát

Chúng tôi thực hiện thử nghiệm trên nhiều nhiệm vụ đư ợc giám sát bao gồm suy luận ngôn ngữ tự nhiên, trả lời câu hỏi, tư ơ ng đồng về ngữ nghĩa và phân loại văn bản. Một số tác vụ này có sẵn như một phần của điểm chuẩn đa tác vụ GLUE đư ợc phát hành gần đây [64] mà chúng tôi sử dụng. Hình 1 cung cấp tổng quan về tất cả các nhiệm vụ và bộ dữ liệu.

Suy luận bằng ngôn ngữ tự nhiên Nhiệm vụ của suy luận bằng ngôn ngữ tự nhiên (NLI), còn đư ợc gọi là nhận diện sự nối tiếp trong văn bản, liên quan đến việc đọc một cặp câu và đánh giá mối quan hệ giữa chúng theo một trong những câu nối tiếp, mâu thuẫn hoặc trung lập. Mặc dù gần đây đã có rất nhiều sự quan tâm [58, 35, 44], như ng nhiệm vụ này vẫn còn nhiều thách thức do sự hiện diện của nhiều hiện tư ợng như sự kéo theo từ vựng, sự đồng quy, và sự mơ hồ về từ vựng và cú pháp. Chúng tôi đánh giá trên năm bộ dữ liệu với nhiều nguồn khác nhau, bao gồm chú thích hình ảnh (SNLI), bài phát biểu đư ợc phiên âm, tiểu thuyết phổ biến và báo cáo của chính phủ (MNLI), bài báo Wikipedia (QNLI), bài kiểm tra khoa học (SciTail) hoặc bài báo (RTE).

Bảng 2 nêu chi tiết các kết quả khác nhau về các nhiệm vụ NLI khác nhau đối với mô hình của chúng tôi và các phư ơ ng pháp tiếp cận hiện đại trư ớc đây . Phư ơ ng pháp của chúng tôi vư ợt trội hơ n đáng kể so với đư ờng cơ sở trên bốn trong số năm bộ dữ liệu, đạt đư ợc mức cải thiện tuyệt đối lên tới 1,5% trên MNLI, 5% trên SciTail, 5,8% trên QNLI và 0,6% trên SNLI so với kết quả tốt nhất trư ớc đó. Điều này chứng tỏ mô hình của chúng tôi có khả năng suy luận tốt hơ n đối với nhiều câu và xử lý các khía cạnh của sự mơ hồ về ngôn ngữ. Trên RTE, một trong những bộ dữ liệu nhỏ hơ n mà chúng tôi đánh giá (2490 ví dụ), chúng tôi đạt đư ợc độ chính xác là 56%, thấp hơ n mức 61,7% đư ợc báo cáo bởi mô hình biLSTM đa tác vụ. Với hiệu suất mạnh mẽ của phư ơ ng pháp tiếp cận của chúng tôi trên các bộ dữ liệu NLI lớn hơ n, có khả năng mô hình của chúng tôi cũng sẽ đư ợc hư ởng lợi từ việc đào tạo đa tác vụ như ng hiện tại chúng tôi chư a khám phá điều này.

[2] https://ftfy.readthedocs.io/en/latest/
[3] https://spacy.io/

Bảng 2: Kết quả thực nghiệm về các tác vụ suy luận ngôn ngữ tự nhiên, so sánh mô hình của chúng tôi với mô hình hiện tại các phư ơ ng pháp tiên tiến nhất. 5x biểu thị một tập hợp gồm 5 mô hình. Tất cả các bộ dữ liệu sử dụng độ chính xác như thư ớc đo đánh giá.

Phư ơ ng pháp	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89,3	-	-	-
CÀ PHÊ [58] (5x)	80,2	79,0	89,3	-	-	-
Mạng trả lời ngẫu nhiên [35] (3x)	80,6	80,1	-	-	-	-
CÀ PHÊ [58]	78,7	77,9	88,5	83.3		
GenSen [64]	71,4	71,3	-	-	82,3	59,2
Đa tác vụ BiLSTM + Attn [64]	72,2	72,1	-	-	82,1	61,7
Máy biến áp tinh chỉnh LM (của chúng tôi)	82.1	81,4	89,9	88.3	88.1	56.0

Bảng 3: Kết quả trả lời câu hỏi và suy luận hợp lý, so sánh mô hình của chúng tôi với các phư ơ ng pháp tiên tiến nhất hiện nay.. 9x có nghĩa là một tập hợp gồm 9 mô hình.

Phư ơ ng pháp	Câu chuyện Cloze	RACE-m	RACE-h	RACE
val-LS-bỏ qua [55]	76,5	-	-	-
Mô hình mạch lạc ẩn [7]	77,6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55,6	49,4	51.2
BiAttention MRU [59] (9x)	-	60,2	50,3	53.3
Máy biến áp tinh chỉnh LM (của chúng tôi)	86,5	62,9	57,4	59,0

Trả lời câu hỏi và suy luận hợp lý Một nhiệm vụ khác đòi hỏi các khía cạnh của một và lập luận nhiều câu là trả lời câu hỏi. Chúng tôi sử dụng bộ dữ liệu RACE đư ợc phát hành gần đây [30], bao gồm các đoạn tiếng Anh với các câu hỏi liên quan từ các kỳ thi trung học cơ sở và trung học phổ thông. Cái này kho văn bản đã đư ợc chứng minh là chứa nhiều câu hỏi loại lý luận hơ n các bộ dữ liệu khác như CNN [19] hoặc SQuaD [47], cung cấp đánh giá hoàn hảo cho mô hình của chúng tôi đư ợc đào tạo để xử lý tầm xa bối cảnh. Ngoài ra, chúng tôi đánh giá trên Bài kiểm tra Story Cloze [40], bao gồm việc chọn đúng kết thúc cho câu chuyện nhiều câu từ hai phư ơ ng án. Trong các nhiệm vụ này, mô hình của chúng tôi lại vư ợt trội so với mô hình kết quả tốt nhất trư ớc đó với tỷ suất lợi nhuận đáng kể - lên tới 8,9% trên Story Cloze và 5,7% tổng thể trên RACE. Điều này chứng tỏ khả năng mô hình của chúng tôi xử lý các bối cảnh tầm xa một cách hiệu quả.

Tư ơ ng tự ngữ nghĩa Các nhiệm vụ tư ơ ng tự ngữ nghĩa (hoặc phát hiện diễn giải) liên quan đến việc dự đoán liệu hai câu có tư ơ ng đư ơ ng về mặt ngữ nghĩa hay không. Những thách thức nằm trong việc nhận ra diễn đạt lại của khái niệm, hiểu phủ định, và xử lý sự mơ hồ cú pháp. Chúng tôi sử dụng ba bộ dữ liệu cho việc này nhiệm vụ – Microsoft Paraphrase corpus (MRPC) [14] (thu thập từ các nguồn tin tức), Quora Bộ dữ liệu Cặp câu hỏi (QQP) [9] và điểm chuẩn Tư ơ ng tự văn bản ngữ nghĩa (STS-B) [6]. Chúng tôi thu đư ợc kết quả tiên tiến nhất về hai trong số ba nhiệm vụ tư ơ ng tự về ngữ nghĩa (Bảng 4) với 1 tăng điểm tuyệt đối trên STS-B. Đồng bằng hiệu suất trên QQP là đáng kể, với 4,2% tuyệt đối cải tiến so với BiLSTM + ELMo + Attn một nhiệm vụ.

Phân loại Cuối cùng, chúng tôi cũng đánh giá trên hai nhiệm vụ phân loại văn bản khác nhau. Tập văn bản về khả năng chấp nhận ngôn ngữ (CoLA) [65] chứa các đánh giá của chuyên gia về việc liệu một câu có ngữ pháp hay không, và kiểm tra khuynh hư ớng ngôn ngữ bẩm sinh của các mô hình đư ợc đào tạo. Tình cảm Stanford Mặt khác, Treebank (SST-2) [54] là một nhiệm vụ phân loại nhị phân tiêu chuẩn. Mô hình của chúng tôi thu đư ợc điểm số 45,4 trên CoLA, đây là một bư ớc nhảy đặc biệt lớn so với kết quả tốt nhất trư ớc đó là 35,0, thể hiện sự thiên vị ngôn ngữ bẩm sinh đư ợc học bởi mô hình của chúng tôi. Mô hình cũng đạt độ chính xác 91,3% on SST-2, co the cạnh tranh với các kết quả hiện đại. Chúng tôi cũng đạt đư ợc tổng điểm là 72,8 trên điểm chuẩn GLUE, tốt hơ n đáng kể so với mức tốt nhất trư ớc đó là 68,9.

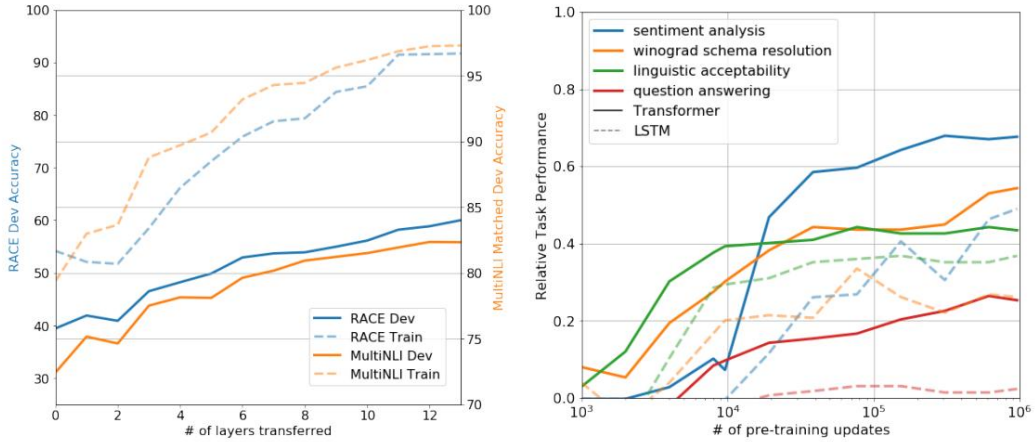
Bảng 4: Kết quả phân loại và tư ơ ng đồng về ngữ nghĩa, so sánh mô hình của chúng tôi với các phư ơ ng pháp tiên tiến nhất hiện nay. Tất cả các đánh giá nhiệm vụ trong bảng này đã đư ợc thực hiện bằng điểm chuẩn GLUE. (mc=Mathews tư ơ ng quan, acc=Độ chính xác, pc=tư ơ ng quan Pearson)

Phư ơ ng pháp	Phân loại	Tư ơ ng tự ngữ nghĩa GLUE				
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (máy tính)	QQP (F1)	
mLSTM byte thư a thớt [16]	-	93.2	-	-	-	-
TF-KLD [23]	-	-	86,0	-	-	-
ECNU (nhóm hỗn hợp) [60]	-	-	-	81,0	-	-
BiLSTM đơ n nhiệm vụ + ELMo + Attn [64] 35.0 BiLSTM		90,2	80,2	55,5	66,1	64,8
đa nhiệm vụ + ELMo + Attn [64] 18.9 Máy biến áp tinh		91,6	83,5	72,8	63,3	68,9
chỉnh LM (của chúng tôi)	45,4	91.3	82.3	82,0	70,3	72,8

Nhìn chung, cách tiếp cận của chúng tôi đạt đư ợc kết quả mới nhất ở 9 trong số 12 bộ dữ liệu mà chúng tôi đánh giá tiếp tục, vư ợt trội so với các nhóm trong nhiều trư ờng hợp. Kết quả của chúng tôi cũng chỉ ra rằng cách tiếp cận của chúng tôi hoạt động tốt trên các bộ dữ liệu có kích thư ớc khác nhau, từ các bộ dữ liệu nhỏ hơ n như STS-B (ví dụ đào tạo ≈5,7k) – đến cái lớn nhất – SNLI (ví dụ đào tạo ≈550k).

5 Phân tích

Tác động của số lớp đư ợc chuyển Chúng tôi quan sát thấy tác động của việc chuyển một số lư ợng thay đổi của các lớp từ đào tạo trư ớc không giám sát đến nhiệm vụ mục tiêu đư ợc giám sát. Hình 2 (trái) minh họa hiệu suất của phư ơ ng pháp của chúng tôi trên MultiNLI và RACE như là một chức năng của số lư ợng lớp đư ợc chuyển. Chúng tôi quan sát thấy kết quả tiêu chuẩn rằng việc chuyển nhúng sẽ cải thiện hiệu suất và mỗi lớp máy biến áp cung cấp thêm lợi ích lên tới 9% khi truyền toàn bộ trên MultiNLI. Điều này chỉ ra rằng mỗi lớp trong mô hình đư ợc đào tạo trư ớc chứa chức năng hữu ích để giải quyết các nhiệm vụ mục tiêu.



Hình 2: (trái) Hiệu ứng chuyển số lư ợng lớp ngày càng tăng từ ngôn ngữ đư ợc đào tạo trư ớc mô hình trên RACE và MultiNLI. (phải) Biểu đồ thể hiện sự phát triển của hiệu suất zero-shot trên các nhiệm vụ khác nhau như một chức năng của các bản cập nhật trư ớc khi đào tạo LM. Hiệu suất trên mỗi tác vụ đư ợc chuẩn hóa giữa đư ờng cơ sở dự đoán ngẫu nhiên và công nghệ tiên tiến nhất hiện nay với một mô hình duy nhất.

Zero-shot Behaviors Chúng tôi muốn hiểu rõ hơ n tại sao việc đào tạo trư ớc mô hình ngôn ngữ cho các biến hình lại hiệu quả. Một giả thuyết là mô hình tổng quát cơ bản học cách thực hiện nhiều các nhiệm vụ chúng tôi đánh giá để cải thiện khả năng lập mô hình ngôn ngữ của nó và càng có cấu trúc

Bảng 5: Phân tích các loại bỏ mô hình khác nhau trên các nhiệm vụ khác nhau. Trung bình điểm là điểm trung bình không trọng số của tất cả các kết quả. (mc= Tư ơ ng quan Mathews, acc=Độ chính xác, pc=Tư ơ ng quan Pearson)

Phư ơ ng pháp	Trung bình Điểm CoLA SST2 MRPC STSB QQP MNLI QNLI RTE								
		(mc)	(acc)	(F1)	(máy tính)	(F1)	(acc)	(acc)	(acc)
Biến áp có aux LM (đầy đủ)	74,7	45,4	91.3	82.3	82,0	70.3	81,8	88.1	56,0
Máy biến áp không có đào tạo trư ớc	59,9	18,9	84,0	79,4	30,9	65,5	75,7	71,2	53,8
Máy biến áp không có phụ trợ LM	75,0	47,9	92,0	84,9	83,2	69,8	81,1	86,9	54,4
LSTM với phụ trợ LM	69,1	30,3	90,5	83,2	71,8	68,1	73,7	81,1	54,6

bộ nhớ chú ý của máy biến áp hỗ trợ chuyển giao so với LSTM. Chúng tôi đã thiết kế một loạt của các giải pháp heuristic sử dụng mô hình tổng quát cơ bản để thực hiện các nhiệm vụ mà không cần giám sát tinh chỉnh. Chúng tôi hình dung hiệu quả của các giải pháp heuristic này trong quá trình tạo ra đào tạo trư ớc trong Hình 2 (phải). Chúng tôi quan sát hiệu suất của các heuristic này là ổn định và đều đặn tăng trong quá trình đào tạo cho thấy rằng đào tạo trư ớc tổng quát hỗ trợ việc học nhiều loại của chức năng liên quan đến nhiệm vụ. Chúng tôi cũng quan sát thấy LSTM thể hiện phư ơ ng sai cao hơ n trong cú đánh không của nó hiệu suất cho thấy rằng xu hư ớng quy nạp của kiến trúc Máy biến áp hỗ trợ chuyển giao.

Đối với CoLA (khả năng chấp nhận ngôn ngữ), các ví dụ đư ợc chấm điểm dư ới dạng xác suất nhật ký mã thông báo trung bình mô hình tổng quát chỉ định và dự đoán đư ợc thực hiện bằng ngư ững. Đối với SST-2 (phân tích tình cảm), chúng tôi thêm mã thông báo vào từng ví dụ và chỉ giới hạn phân phối đầu ra của mô hình ngôn ngữ các từ tích cực và tiêu cực và đoán mã thông báo mà nó chỉ định xác suất cao hơ n làm dự đoán. Đối với RACE (trả lời câu hỏi), chúng tôi chọn câu trả lời mà mô hình tổng quát chỉ định mức trung bình cao nhất xác suất đăng nhập mã thông báo khi có điều kiện trên tài liệu và câu hỏi. Đối với CHDCND Triều Tiên [46] (winograd lư ợc đồ), chúng tôi thay thế đại từ xác định bằng hai giới thiệu có thể và dự đoán độ phân giải rằng mô hình tổng quát chỉ định xác suất nhật ký mã thông báo trung bình cao hơ n cho phần còn lại của chuỗi sau khi thay thế.

Nghiên cứu cắt bỏ Chúng tôi thực hiện ba nghiên cứu cắt bỏ khác nhau (Bảng 5). Đầu tiên, chúng tôi kiểm tra các hiệu suất của phư ơ ng pháp của chúng tôi mà không có mục tiêu LM phụ trợ trong quá trình tinh chỉnh. Chúng tôi quan sát thấy rằng mục tiêu phụ giúp thực hiện các nhiệm vụ NLI và QQP. Nhìn chung, xu hư ớng cho thấy rằng các bộ dữ liệu lớn hơ n đư ợc hư ởng lợi từ mục tiêu phụ như ng các bộ dữ liệu nhỏ hơ n thì không. Thứ hai, chúng tôi phân tích ảnh hư ởng của Transformer bằng cách so sánh nó với LSTM đơ n vị 2048 lớp đơ n sử dụng cùng một khung. Chúng tôi quan sát điểm số trung bình giảm 5,6 khi sử dụng LSTM thay vì Transformer. Chỉ LSTM vư ợt trội hơ n Transformer trên một tập dữ liệu – MRPC. Cuối cùng, chúng tôi cũng so sánh với máy biến áp của chúng tôi kiến trúc đư ợc đào tạo trực tiếp trên các nhiệm vụ mục tiêu đư ợc giám sát mà không cần đào tạo trư ớc. Chúng tôi quan sát thấy rằng sự thiếu của đào tạo trư ớc làm ảnh hư ởng đến hiệu suất trên tất cả các nhiệm vụ, dẫn đến giảm 14,8% so với của chúng tôi đầy đủ mô hình.

6 Kết luận

Chúng tôi đã giới thiệu một khuôn khổ để đạt đư ợc sự hiểu biết sâu sắc về ngôn ngữ tự nhiên chỉ bằng một mô hình bất khả tri nhiệm vụ thông qua đào tạo trư ớc tổng quát và tinh chỉnh phân biệt đối xử. Bằng cách đào tạo trư ớc trên một kho văn bản đa dạng với các đoạn văn bản dài liền kề, mô hình của chúng tôi có đư ợc thế giới quan trọng kiến thức và khả năng xử lý các phụ thuộc tầm xa mà sau đó đư ợc chuyển giao thành công cho giải quyết các nhiệm vụ phân biệt như trả lời câu hỏi, đánh giá sự tư ơ ng đồng về ngữ nghĩa, sự đòi hỏi xác định và phân loại văn bản, cải thiện trạng thái của nghệ thuật trên 9 trong số 12 bộ dữ liệu chúng tôi học. Sử dụng đào tạo (trư ớc) không giám sát để tăng hiệu suất cho các nhiệm vụ phân biệt đối xử từ lâu là một mục tiêu quan trọng của nghiên cứu Machine Learning. Công việc của chúng tôi cho thấy rằng việc đạt đư ợc ý nghĩa thực sự có thể đạt đư ợc hiệu suất và đư a ra gợi ý về những kiểu máy (Máy biến áp) và bộ dữ liệu nào (văn bản có phụ thuộc phạm vi dài) hoạt động tốt nhất với phư ơ ng pháp này. Chúng tôi hy vọng rằng điều này sẽ giúp cho phép nghiên cứu mới về học tập không giám sát, cho cả hiểu ngôn ngữ tự nhiên và các lĩnh vực khác, nâng cao hơ n nữa sự hiểu biết của chúng tôi về cách thức và thời điểm học tập không giám sát hoạt động.

Ngư ời giới thiệu

[1] S. Arora, Y. Liang, và T. Ma. Một đư ờng cơ sở đơ n giản như ng khó đánh bại để nhúng câu. 2016.

[2] JL Ba, JR Kiros, và GE Hinton. Chuẩn hóa lớp. bản in trư ớc arXiv arXiv:1607.06450, 2016.

[3] Y. Bengio, P. Lamblin, D. Popovici, và H. Larochelle. Đào tạo theo lớp tham lam của các mạng sâu. TRONG
Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 153–160, 2007.

[4] L. Bentivogli, P. Clark, I. Dagan, và D. Giampiccolo. Pascal thứ năm công nhận văn bản đòi hỏi
thử thách. Trong TAC, 2009.

[5] SR Bowman, G. Angeli, C. Potts và CD Manning. Một kho ngữ liệu lớn đư ợc chú thích để học tự nhiên
suy luận ngôn ngữ. EMNLP, 2015.

[6] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, và L. Specia. Nhiệm vụ Semeval-2017 1: Đánh giá tập trung vào sự
tư ơ ng đồng về ngữ nghĩa của văn bản -đa ngôn ngữ và đa ngôn ngữ. bản in trư ớc arXiv arXiv:1708.00055, 2017.

[7] S. Chaturvedi, H. Peng, và D. Roth. Hiểu câu chuyện để dự đoán những gì xảy ra tiếp theo. Trong Kỷ yếu Hội nghị
2017 về Phư ơ ng pháp Thực nghiệm trong Xử lý Ngôn ngữ Tự nhiên, trang 1603–1614, 2017.

[8] D. Chen và C. Manning. Trình phân tích cú pháp phụ thuộc nhanh và chính xác sử dụng mạng thần kinh. Trong Kỷ yếu
hội nghị 2014 về các phư ơ ng pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên (EMNLP), trang 740–750, 2014.

[9] Z. Chen, H. Zhang, X. Zhang và L. Zhao. cặp câu hỏi Quora. https://data.quora.com/First-Quora
Bộ dữ liệu-Phát hành-Câu hỏi-Cặp, 2018.

[10] R. Collobert và J. Weston. Một kiến trúc hợp nhất để xử lý ngôn ngữ tự nhiên: Mạng thần kinh sâu với khả năng
học tập đa nhiệm. Trong Kỷ yếu hội nghị quốc tế lần thứ 25 về Máy học, trang 160–167. ACCM, 2008.

[11] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, và P. Kuksa. Xử lý ngôn ngữ tự nhiên (gần như)
từ đầu. Tạp chí Nghiên cứu Máy học, 12(Aug):2493–2537, 2011.

[12] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, và A. Bordes. Học có giám sát của câu phổ quát
biểu diễn từ dữ liệu suy luận ngôn ngữ tự nhiên. EMNLP, 2017.

[13] AM Đại và QV Lê. Học trình tự bán giám sát. Những tiến bộ trong xử lý thông tin thần kinh
Hệ thống, trang 3079–3087, 2015.

[14] WB Dolan và C. Rockett. Tự động xây dựng một tập hợp các diễn giải câu cảm. trong thủ tục tố tụng
của Hội thảo quốc tế lần thứ ba về diễn giải (IWP2005), 2005.

[15] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent và S. Bengio. Tại sao đào tạo trư ớc không giám sát
giúp học sâu? Tạp chí Nghiên cứu Máy học, 11(Feb):625–660, 2010.

[16] S. Gray, A. Radford, và KP Diederik. Hạt nhân GPU cho trọng lư ợng khối thư a thớt. 2017.

[17] Z. He, S. Liu, M. Li, M. Zhou, L. Zhang và H. Wang. Biểu diễn thực thể học để phân biệt đối tư ợng thực thể .
Trong Kỷ yếu Hội nghị Thư ờng niên lần thứ 51 của Hiệp hội Ngôn ngữ học Tính toán (Tập 2: Bài viết ngắn), tập
2, trang 30–34, 2013.

[18] D. Hendrycks và K. Gimpel. Cầu nối phi tuyến tính và bộ điều chỉnh ngẫu nhiên với các đơ n vị tuyến tính lỗi
gaussian. bản in trư ớc arXiv arXiv:1606.08415, 2016.

[19] KM Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman và P. Blunsom. Máy dạy đọc và hiểu.
Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, trang 1693–1701 , 2015.

[20] GE Hinton, S. Osindero, và Y.-W. Té. Một thuật toán học nhanh cho mạng lư ới niềm tin sâu sắc. thần kinh
tính toán, 18(7):1527–1554, 2006.

[21] J. Howard và S. Ruder. Tinh chỉnh mô hình ngôn ngữ chung để phân loại văn bản. Hiệp hội cho
Ngôn ngữ học tính toán (ACL), 2018.

[22] Y. Jernite, SR Bowman, và D. Sontag. Các mục tiêu dựa trên diễn ngôn để học biểu diễn câu nhanh không giám sát .
bản in trư ớc arXiv arXiv:1705.00557, 2017.

[23] Y. Ji và J. Eisenstein. Cải tiến phân biệt đối với sự giống nhau của câu phân phối. Trong Kỷ yếu Hội nghị 2013
về Phư ơ ng pháp Thực nghiệm trong Xử lý Ngôn ngữ Tự nhiên, trang 891–896, 2013.

[24] F. Jiao, S. Wang, C.-H. Lee, R. Greiner và D. Schuurmans. Các trường ngẫu nhiên có điều kiện bán giám sát để phân đoạn và ghi nhãn trình tự được cải thiện. Trong Kỷ yếu của Hội nghị Quốc tế về Ngôn ngữ học Tính toán lần thứ 21 và cuộc họp thường niên lần thứ 44 của Hiệp hội Ngôn ngữ học Tính toán, trang 209–216. Hiệp hội Ngôn ngữ học tính toán, 2006.

[25] T. Khot, A. Sabharwal, và P. Clark. Scitail: Một bộ dữ liệu văn bản liên quan đến trả lời câu hỏi khoa học. Trong Kỷ yếu của AAAI, 2018.

[26] Y. Kim. Mạng thần kinh tích chập để phân loại câu. EMNLP, 2014.

[27] DP Kingma và J. Ba. Adam: Một phương pháp tối ưu hóa ngẫu nhiên. arXiv in sẵn arXiv:1412.6980, 2014.

[28] R. Kiros, Y. Zhu, R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, và S. Fidler. Vectơ bỏ qua suy nghĩ . Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 3294–3302, 2015.

[29] N. Kitaev và D. Klein. Phân tích cú pháp khu vực bầu cử bằng bộ mã hóa tự chú ý. ACL, 2018.

[30] G. Lai, Q. Xie, H. Liu, Y. Yang, và E. Hovy. Chủng tộc: Bộ dữ liệu đọc hiểu quy mô lớn từ các kỳ thi. EMNLP, 2017.

[31] G. Lample, L. Denoyer, và M. Ranzato. Dịch máy không giám sát sử dụng kho ngữ liệu đơn ngữ chỉ một. ICLR, 2018.

[32] Q. Lê và T. Mikolov. Biểu diễn phân tán của câu và tài liệu. Trong Hội nghị quốc tế về học máy, trang 1188–1196, 2014.

[33] P. Lương. Học bán giám sát đối với ngôn ngữ tự nhiên. Luận án Tiến sĩ, Viện Công nghệ Massachusetts, 2005.

[34] PJ Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, và N. Shazeer. Tạo wikipedia bằng cách tóm tắt các chuỗi dài. ICLR, 2018.

[35] X. Liu, K. Duh, và J. Gao. Mạng trả lời ngẫu nhiên cho suy luận ngôn ngữ tự nhiên. bản in trước arXiv arXiv:1804.07888, 2018.

[36] L. Logeswaran và H. Lee. Một khuôn khổ hiệu quả để học biểu diễn câu. ICLR, 2018.

[37] I. Loshchilov và F. Hutter. Sửa lỗi phân rã trọng lượng chính quy trong adam. arXiv in sẵn arXiv:1711.05101, 2017.

[38] B. McCann, J. Bradbury, C. Xiong, và R. Socher. Đã học trong bản dịch: Vectơ từ theo ngữ cảnh. Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, trang 6297–6308, 2017.

[39] T. Mikolov, I. Sutskever, K. Chen, GS Corrado, và J. Dean. Biểu diễn phân tán của các từ và cụm từ và thành phần của chúng. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 3111–3119, 2013.

[40] N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, và J. Allen. Nhiệm vụ được chia sẻ của Lsdsem 2017: Bài kiểm tra đóng băng câu chuyện. Trong Kỷ yếu Hội thảo lần thứ 2 về Các mô hình liên kết ngữ nghĩa cấp từ vựng, câu và diễn ngôn, trang 46–51, 2017.

[41] K. Nigam, A. McCallum, và T. Mitchell. Phân loại văn bản bán giám sát sử dụng em. Học bán giám sát , trang 33–56, 2006.

[42] J. Pennington, R. Socher, và C. Manning. Găng tay: Các vectơ tổng thể để biểu diễn từ. Trong Kỷ yếu hội nghị 2014 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên (EMNLP), trang 1532–1543, 2014.

[43] TÔI Peters, W. Ammar, C. Bhagavatula, và R. Power. Gắn thẻ trình tự bán giám sát với bidirec các mô hình ngôn ngữ cụ thể. ACL, 2017.

[44] TÔI Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee và L. Zettlemoyer. Biểu diễn từ được mã hóa theo ngữ cảnh sâu . NAACL, 2018.

[45] Y. Qi, DS Sachan, M. Felix, SJ Padmanabhan, và G. Neubig. Khi nào và tại sao nhúng từ được đào tạo trước lại hữu ích cho dịch máy thần kinh? NAACL, 2018.

[46] A. Rahman và V. Ng. Giải quyết các trường hợp phức tạp của đại từ xác định: thử thách lược đồ winograd. Trong Kỷ yếu của Hội nghị chung năm 2012 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên và học ngôn ngữ tự nhiên tính toán, trang 777–789. Hiệp hội Ngôn ngữ học Tính toán, 2012.

[47] P. Rajpurkar, J. Zhang, K. Lopyrev, và P. Liang. Biệt đội: Hơn 100.000 câu hỏi để máy hiểu văn bản. EMNLP, 2016.

[48] P. Ramachandran, PJ Liu, và QV Le. Đào tạo trước không giám sát cho việc học theo trình tự. bản in trước arXiv arXiv:1611.02683, 2016.

[49] M. Ranzato, C. Poultney, S. Chopra, và Y. LeCun. Học hiệu quả các biểu diễn thưa thớt với mô hình dựa trên năng lượng. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 1137–1144, 2007.

[50] Ông Rei. Học đa nhiệm bán giám sát để ghi nhãn trình tự. ACL, 2017.

[51] H. Robbins và S. Monro. Một phương pháp xấp xỉ ngẫu nhiên. Biên niên sử thống kê toán học, trang 400–407, 1951.

[52] T. Rocktäschel, E. Grefenstette, KM Hermann, T. Kocisky, và P. Blunsom. Lý luận về sự đòi hỏi với sự chú ý thần kinh. bản in trước arXiv arXiv:1509.06664, 2015.

[53] R. Sennrich, B. Haddow, và A. Birch. Dịch máy thần kinh của các từ hiếm với các đơn vị từ phụ. arXiv bản in trước arXiv:1508.07909, 2015.

[54] R. Socher, A. Perelygin, J. Wu, J. Chuang, CD Manning, A. Ng, và C. Potts. Các mô hình sâu đệ quy cho thành phần ngữ nghĩa trên một treebank tình cảm. Trong Kỷ yếu hội nghị 2013 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên, trang 1631–1642, 2013.

[55] S. Srinivasan, R. Arora, và M. Riedl. Một cách tiếp cận đơn giản và hiệu quả để kiểm tra câu chuyện cloze. arXiv bản in trước arXiv:1803.05547, 2018.

[56] S. Subramanian, A. Trischler, Y. Bengio, và CJ Pal. Học các biểu diễn câu phân tán cho mục đích chung thông qua học tập đa tác vụ quy mô lớn. bản in trước arXiv arXiv:1804.00079, 2018.

[57] J. Suzuki và H. Isozaki. Ghi nhãn tuần tự bán giám sát và phân đoạn sử dụng thang giga-word dữ liệu chưa gán nhãn. Kỷ yếu ACL-08: HLT, trang 665–673, 2008.

[58] Y. Tay, LA Tuan, và SC Hui. Kiến trúc so sánh lan truyền với hệ số căn chỉnh cho suy luận ngôn ngữ tự nhiên. bản in trước arXiv arXiv:1801.00102, 2017.

[59] Y. Tay, LA Tuan, và SC Hui. Lý luận đa phạm vi để hiểu máy. bản in trước arXiv arXiv:1803.09074, 2018.

[60] J. Tian, Z. Zhou, M. Lan, và Y. Wu. Ecnu tại nhiệm vụ 1 trong học kỳ 2017: Tận dụng các tính năng nlp truyền thống dựa trên hạt nhân và mạng lưới thần kinh để xây dựng một mô hình phổ quát cho sự tương đồng về văn bản ngữ nghĩa đa ngôn ngữ và đa ngôn ngữ. Trong Kỷ yếu Hội thảo quốc tế về đánh giá ngữ nghĩa lần thứ 11 (SemEval-2017), trang 191–197, 2017.

[61] Y. Tsvetkov. Cơ hội và thách thức khi làm việc với các ngôn ngữ tài nguyên thấp. CMU, 2017.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN Gomez, Ł. Kaiser và I. Polosukhin. Chú ý là tất cả những gì bạn cần. Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, trang 6000–6010, 2017.

[63] P. Vincent, H. Larochelle, Y. Bengio, và P.-A. Manzagol. Trích xuất và tổng hợp các tính năng mạnh mẽ với bộ mã hóa tự động khử nhiễu. Trong Kỷ yếu của hội nghị quốc tế lần thứ 25 về Máy học, trang 1096–1103. ACCM, 2008.

[64] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, và SR Bowman. Glue: Nền tảng phân tích và điểm chuẩn đa tác vụ để hiểu ngôn ngữ tự nhiên. bản in trước arXiv arXiv:1804.07461, 2018.

[65] A. Warstadt, A. Singh, và SR Bowman. Corpus của khả năng chấp nhận ngôn ngữ. http://nyu-mll.github.io/cola, 2018.

[66] A. Williams, N. Nangia, và SR Bowman. Kho ngữ liệu thách thức bao quát để hiểu câu thông qua suy luận. NAACL, 2018.

[67] Y. Xu, J. Liu, J. Gao, Y. Shen, và X. Liu. Hướng tới khả năng đọc hiểu máy ở cấp độ con người: Lập luận và suy luận bằng nhiều chiến lược. bản in trước arXiv arXiv:1711.04964, 2017.

[68] D. Yu, L. Deng, và G. Dahl. Vai trò của đào tạo trư ớc và tinh chỉnh trong dbn-hmms phụ thuộc vào ngữ cảnh để nhận dạng giọng nói trong thế giới thực. Ở Proc. Hội thảo NIPS về Deep Learning và Unsupervised Feature Learning, 2010.

[69] R. Zhang, P. Isola, và AA Efros. Bộ mã hóa tự động chia não: Học không giám sát bằng dự đoán kênh chéo . Trong CVPR, tập 1, trang 6, 2017.

[70] X.Trư . Khảo sát văn học học bán giám sát. 2005.

[71] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, và S. Fidler. Sắp xếp sách và phim: Hư ớng tới giải thích trực quan giống như câu chuyện bằng cách xem phim và đọc sách. Trong Kỷ yếu của hội nghị quốc tế IEEE về thị giác máy tính, trang 19–27, 2015.