Explainable Outfit Recommendation with Joint Outfit Matching and Comment Generation

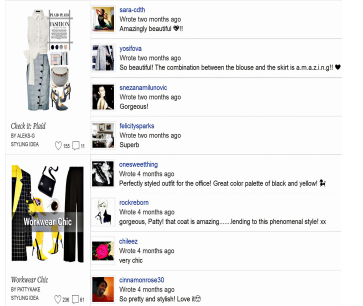
(KHUYẾN NGHỊ TRANG PHỤC CÓ GIẢI THÍCH VỚI KẾT HỢP TRANG PHỤC VÀ TẠO BÌNH LUẬN)

Tóm tắt - Hầu hết các công việc trước đây về đề xuất trang phục tập trung vào việc thiết kế các tính năng trực quan để tăng cường các đề xuất. Công việc hiện tại bỏ qua ý kiến ​​của người dùng về các mặt hàng thời trang, đã được chứng minh là có hiệu quả trong việc tạo ra các giải thích cùng với kết quả đề xuất tốt hơn. Chúng tôi đề xuất một khung mạng thần kinh mới, khuyến nghị trang phục thần kinh (NOR), đồng thời cung cấp các khuyến nghị về trang phục và tạo ra các nhận xét trừu tượng. NOR bao gồm hai phần: trang phục phù hợp và tạo bình luận. Để phù hợp với trang phục, chúng tôi đề xuất một mạng lưới thần kinh tích chập với cơ chế chú ý lẫn nhau để trích xuất các đặc điểm hình ảnh. Các tính năng hình ảnh sau đó được giải mã thành một điểm đánh giá cho dự đoán phù hợp. Để tạo bình luận trừu tượng, chúng tôi đề xuất một mạng lưới thần kinh tái phát có kiểm soát với cơ chế chú ý đa phương thức để chuyển đổi các đặc điểm hình ảnh thành một câu ngắn gọn. Hai phần được đào tạo chung dựa trên khung học tập đa tác vụ trong mô hình truyền bá ngược từ đầu đến cuối. Các thí nghiệm mở rộng được thực hiện trên một bộ dữ liệu hiện có và một bộ dữ liệu trong thế giới thực được thu thập cho thấy NOR đạt được những cải tiến đáng kể so với các đường cơ sở hiện đại để đề xuất trang phục. Trong khi đó, các bình luận được tạo ra của chúng tôi đạt được điểm ROUGE và BLEU ấn tượng so với các bình luận do con người viết. Các ý kiến ​​được tạo ra có thể được coi là giải thích cho kết quả đề xuất. Chúng tôi phát hành tập dữ liệu và mã để tạo điều kiện cho nghiên cứu trong tương lai.

Index Terms—Outfit recommendation, explainable recommendation

1. **GIỚI THIỆU**

khuyến nghị ouffit đóng một vai trò ngày càng quan trọng trong thị trường bán lẻ trực tuyến. Mục đích của đề xuất trang phục là để thúc đẩy sự quan tâm và tham gia của mọi người vào mua sắm trực tuyến bằng cách đề xuất trang phục thời trang mà họ có thể quan tâm. Các nghiên cứu ban đầu về khuyến nghị trang phục dựa trên các bộ dữ liệu nhỏ nhưng có chú thích của chuyên gia [1, 2], ngăn cấm sự phát triển của các mô hình phức tạp cần các bộ tài liệu đào tạo lớn (ví dụ: các mô hình dựa trên học tập sâu). Trong những năm gần đây, với sự phát triển của các mối quan hệ trực tuyến theo định hướng thời trang, ví dụ, Polyvore 2 và Chictopia, 3 người có thể chia sẻ và nhận xét về các thành phần trang phục, như trong Hình 1. Ngoài một số lượng lớn các thành phần trang phục, dữ liệu đám đông như vậy cũng chứa thông tin có giá trị (ví dụ: nhận xét của người dùng) để xây dựng các hệ thống đề xuất thông minh và chính xác hơn



Hình 1: Trang phục có ý kiến người dùng từ Polyvore. Người dùng chia sẻ các thành phần trang phục của họ với công chúng rộng rãi (trái) và một người khác bày tỏ ý kiến của họ về các thành phần trang phục (phải).

Chúng tôi giải quyết các nhiệm vụ khuyến nghị trang phục có giải thích. Đưa ra một phần trên (ví dụ, áo ngoài), chúng tôi cần đề xuất một danh sách ngắn các phần dưới (ví dụ: quần hoặc váy) từ một bộ sưu tập lớn phù hợp nhất với phần trên và trong khi đó tạo ra một câu cho mỗi đề xuất để giải thích lý do tại sao kết hợp phần trên với phần dưới, và ngược lại. Bằng cách giải thích lý do tại sao một bộ trang phục được khuyến nghị, hệ thống đề xuất trở nên minh bạch và đáng tin cậy hơn, giúp người dùng đưa ra quyết định nhanh hơn và tốt hơn [3]. Nhiệm vụ của đề xuất trang phục có thể giải thích là không tầm thường bởi vì có hai vấn đề chính: (1) Chúng ta cần mô hình hóa tính tương thích của các yếu tố thời trang, ví dụ: màu sắc, chất liệu, hoa văn, hình dạng, v.v. [4]. (2) Chúng ta cần mô hình hóa các biến đổi giữa thông tin hình ảnh và thông tin văn bản, bao gồm ánh xạ từ hình ảnh đến không gian văn bản

Để giải quyết các vấn đề được liệt kê ở trên, chúng tôi đề xuất một khung học tập đa tác vụ thần kinh, được gọi là khuyến nghị trang phục thần kinh (NOR). NOR bao gồm hai thành phần cốt lõi: kết hợp trang phục và tạo bình luận. Để kết hợp trang phục, chúng tôi sử dụng mạng lưới thần kinh tích chập (CNN) với cơ chế chú ý lẫn nhau để trích xuất các đặc điểm hình ảnh của trang phục. Cụ thể, trước tiên, chúng tôi sử dụng CNN để mô hình hóa đỉnh và đáy như các vectơ tiềm ẩn; sau đó chúng tôi đề xuất một cơ chế chú ý lẫn nhau, trích xuất các đặc điểm hình ảnh tốt hơn của cả đỉnh và đáy bằng cách sử dụng các vectơ trên để khớp với các vectơ dưới và ngược lại. Các tính năng hình ảnh sau đó được giải mã thành một điểm đánh giá như dự đoán phù hợp. Để tạo bình luận trừu tượng, chúng tôi đề xuất một mạng nơ ron tái phát có kiểm soát (RNN) với cơ chế chú ý đa phương thức để chuyển đổi các đặc điểm hình ảnh thành một câu ngắn gọn. Cụ thể, để tạo một từ, NOR học cách ánh xạ giữa không gian hình ảnh và văn bản, điều này đạt được với cơ chế chú ý đa phương thức. Tất cả các tham số thần kinh trong hai phần của khung của chúng tôi cũng như các từ nhúng được học theo cách tiếp cận học tập đa tác vụ trong mô hình đào tạo truyền bá từ đầu đến cuối.

Đã có một số nghiên cứu về khuyến nghị trang phục [1, 2, 5]. Công việc tương tự như của chúng tôi là của Song et al. [4], người đầu tiên sử dụng mạng mã hóa tự động kép để tìm hiểu không gian tương thích tiềm ẩn, trong đó họ cùng mô hình hóa mối quan hệ kết hợp giữa các tính năng trực quan (nghĩa là hình ảnh) và các tính năng theo ngữ cảnh (nghĩa là, danh mục, thẻ). Sau đó, họ sử dụng xếp hạng cá nhân hóa Bayes (BPR) [6] để khai thác các ưu tiên theo cặp giữa đỉnh và đáy. Sự khác biệt giữa công việc của chúng tôi và của họ: Đầu tiên, mô hình của chúng tôi không chỉ có thể đề xuất đỉnh và đáy, mà còn tạo ra một câu có thể đọc được như một nhận xét. Thứ hai, chúng tôi giới thiệu một cơ chế chú ý lẫn nhau và phương thức chéo vào không gian tương thích tiềm ẩn thay vì mạng tự động mã hóa kép. Thứ ba, chúng tôi cùng nhau đào tạo tính năng trích xuất và xếp hạng ưu tiên trong một sơ đồ truyền bá ngược

Chúng tôi thu thập một bộ dữ liệu thực tế lớn từ Polyvore. Tập dữ liệu của chúng tôi chứa thông tin đa phương thức, ví dụ: hình ảnh, siêu dữ liệu theo ngữ cảnh của các mục và nhận xét của người dùng, v.v. Kết quả thử nghiệm mở rộng được thực hiện trên bộ dữ liệu này cho thấy rằng NOR đạt được hiệu suất tốt hơn so với các mô hình hiện đại về khuyến nghị trang phục, về mặt AUC, MAP và MRR. Hơn nữa, các bình luận được tạo ra từ NOR đạt được điểm ROUGE và BLEU ấn tượng.

Tóm lại, những đóng góp của chúng tôi là:

* Chúng tôi khám phá ý kiến người dùng để cải thiện chất lượng đề xuất trang phục cùng với giải thích
* Chúng tôi đề xuất một khung dựa trên học tập sâu có tên là NOR có thể đồng thời đưa ra các đề xuất về trang phục và tạo ra các nhận xét trừu tượng với chất lượng ngôn ngữ tốt mô phỏng trải nghiệm và cảm xúc của công chúng
* Chúng tôi sử dụng sự chú ý lẫn nhau để mô hình hóa sự tương thích giữa các mặt hàng thời trang và sự chú ý đa phương thức để mô hình hóa sự chuyển đổi giữa không gian hình ảnh và văn bản
* Phương pháp đề xuất của chúng tôi được chứng minh là có hiệu quả trong các thử nghiệm trên bộ dữ liệu hiện có và bộ dữ liệu quy mô lớn được xây dựng có mục đích.

1. **CÔNG VIỆC LIÊN QUAN**

Không có công việc trước đây đã nghiên cứu nhiệm vụ đề xuất trang phục có thể giải thích bằng cách tạo ra các bình luận ngôn ngữ tự nhiên như lời giải thích. Chúng tôi khảo sát ngắn gọn công việc liên quan đến đề xuất trang phục, về đề xuất có thể giải thích và về việc tạo văn bản, tương ứng

* 1. **Khuyến nghị trang phục**

Đưa ra một bức ảnh của một mặt hàng thời trang (ví dụ: top), một hệ thống đề xuất trang phục cố gắng đề xuất một bức ảnh của các mặt hàng thời trang khác (ví dụ: bot). Đã có một số ít nỗ lực để giải quyết nhiệm vụ. Iwata và cộng sự. [1] đề xuất một mô hình chủ đề xác suất để đề xuất đỉnh cho đáy bằng cách tìm hiểu thông tin về tọa độ từ các tính năng trực quan trong từng khu vực mặt hàng thời trang. Liu và cộng sự [2] nghiên cứu cả vấn đề đề nghị trang phục và vật phẩm. Họ đề xuất một mô hình Máy Vector Hỗ trợ tiềm ẩn cho đề xuất trang phục theo định hướng, nghĩa là, trong một dịp đầu vào của người dùng, đề xuất quần áo phù hợp nhất hoặc đề xuất các mặt hàng để kết hợp với quần áo tham khảo. Jagadeesh et al. [7] đề xuất hai lớp người giới thiệu trang phục, cụ thể là xác định và ngẫu nhiên, trong khi họ chủ yếu tập trung vào mô hình màu cho khuyến nghị trang phục.

Các nghiên cứu được liệt kê ở trên hầu hết dựa trên một bộ dữ liệu nhỏ, được chú thích thủ công, ngăn chặn sự phát triển của các mô hình phức tạp. Một số ấn phẩm gần đây đã sử dụng các nguồn khác, trong đó dữ liệu phong phú có thể được thu thập tự động, ví dụ: trong khu vực đề xuất toàn bộ trang phục được cá nhân hóa. Hu et al. [5] đề xuất một phương pháp nhân tố kéo căng chức năng để mô hình hóa các tương tác giữa người dùng và các mặt hàng thời trang qua bộ dữ liệu được thu thập từ Polyvore. McAuley và cộng sự. [8] sử dụng một khung chung để mô hình hóa sở thích trực quan của con người đối với một cặp đối tượng từ bộ dữ liệu đồng mua hàng của Amazon; họ trích xuất các tính năng trực quan với CNN và giới thiệu một số liệu tương tự để khám phá các mối quan hệ trực quan. Tương tự như vậy, He và McAuley [9] giới thiệu một cách tiếp cận nhân tố ma trận kết hợp các tín hiệu thị giác vào các yếu tố dự đoán về ý kiến ​​của người dân. Để xem xét thông tin theo ngữ cảnh (như tiêu đề và danh mục), Li et al. [10] phân loại một bộ trang phục nhất định là phổ biến hoặc không phổ biến thông qua hệ thống học tập sâu đa phương thức và đa thể. Để tổng hợp dữ liệu đa phương thức của các mặt hàng thời trang và thông tin theo ngữ cảnh, Song et al. [4] trước tiên sử dụng một bộ mã hóa tự động để khai thác không gian tương thích tiềm ẩn của chúng. Sau đó, họ sử dụng xếp hạng cá nhân hóa Bayes để khai thác các ưu tiên theo cặp giữa đỉnh và đáy. Kang và cộng sự. [11] sử dụng CNN để tìm hiểu các biểu diễn hình ảnh và cùng sử dụng bộ lọc cộng tác để đề xuất các mặt hàng thời trang cho người dùng. Han và cộng sự. [12] đề xuất để cùng tìm hiểu các nhúng nhúng trực quan và mối quan hệ tương thích giữa các mặt hàng thời trang theo cách từ đầu đến cuối. Họ huấn luyện một mô hình LSTM hai chiều để dự đoán tuần tự các mục tiếp theo dựa trên các mục trước để tìm hiểu các mối quan hệ tương thích của chúng. Song và cộng sự. [13] xem xét kiến ​​thức miền thời trang để kết hợp quần áo và đề xuất một kế hoạch giáo viên để tích hợp nó với các mạng lưới thần kinh. Và họ cũng giới thiệu một sơ đồ chu đáo để gán niềm tin quy tắc trong quy trình chưng cất kiến ​​thức.

Mặc dù ngày càng có nhiều nghiên cứu về đề xuất trang phục, không ai trong số họ đưa ý kiến của người dùng vào tài khoản và không ai có thể đưa ra cả khuyến nghị và nhận xét dễ đọc như chúng tôi làm trong bài viết này.

* 1. **Khuyến nghị giải thích**

Đề xuất có thể giải thích không chỉ cung cấp danh sách các mục được xếp hạng mà còn đưa ra giải thích cho từng mục được đề xuất

Công việc hiện tại về khuyến nghị có thể giải thích có thể được phân loại thành các loại khác nhau, tùy thuộc vào định nghĩa của giải thích được sử dụng. Ở đây, chúng tôi chỉ khảo sát các nghiên cứu liên quan chặt chẽ nhất. Vig et al. [14] đề xuất một phương pháp đề xuất có thể giải thích sử dụng các tag để tạo ra các giải thích. Zhang và cộng sự. [15] đề xuất một mô hình nhân tố rõ ràng để dự đoán xếp hạng trong khi tạo giải thích cấp độ tính năng về lý do tại sao một mặt hàng được hoặc không được đề xuất. Ông và cộng sự. [16] đề xuất TriRank và tích hợp các mô hình chủ đề để tạo ra các yếu tố tiềm ẩn cho người dùng và các mục để đề xuất nhận xét. Ribeiro và cộng sự. [17] đề xuất LIME, một kỹ thuật giải thích mới lạ giải thích các dự đoán của bất kỳ phân loại nào theo cách có thể hiểu và trung thành, bằng cách tìm hiểu một mô hình có thể diễn giải cục bộ xung quanh một dự đoán riêng lẻ. Ren và cộng sự. [18] đề xuất một khái niệm phong phú hơn về giải thích được gọi là quan điểm, được thể hiện dưới dạng một bộ tính năng khái niệm, một chủ đề và nhãn tình cảm; mặc dù họ cung cấp giải thích cho các khuyến nghị, các giải thích là các thẻ đơn giản hoặc các từ hoặc cụm từ được trích xuất. Wang và cộng sự. [19] phát triển một giải pháp học tập đa tác vụ, sử dụng hệ số tenxơ chung để mô hình hóa ưu tiên người dùng cho đề xuất và nội dung có ý kiến ​​để giải thích; thuật toán có thể tạo ra các giải thích bằng cách chiếu các tính năng và cụm từ có liên quan lên không gian được kéo dài bởi các yếu tố người dùng và vật phẩm. Ngược lại, chúng tôi tạo ra các câu ngắn gọn thể hiện lý do tại sao chúng tôi đề xuất một bộ trang phục dựa trên tất cả các nhận xét của người dùng. Và chúng tôi tin rằng việc mô phỏng người dùng để tạo bình luận là cách tốt hơn để gần gũi hơn với quan điểm của người dùng, điều này thể hiện đầy đủ trải nghiệm và cảm xúc của người dùng, giúp người dùng dễ hiểu và chấp nhận hơn

Một số công việc gần đây tạo ra văn bản như giải thích trong khi cung cấp các khuyến nghị. Ni et al. [20] cùng thực hiện đề xuất cá nhân và tạo đánh giá bằng cách kết hợp lọc cộng tác với các mô hình tạo chung dựa trên LSTM. Li et al. [21] Công việc của người tương tự như chúng ta. Bằng cách đưa RNN vào bộ lọc cộng tác, họ cùng nhau dự đoán xếp hạng và tạo mẹo, thể hiện tình cảm của người dùng trong khi xem xét một mục. Công việc của chúng tôi và những người trước đó khác nhau theo bốn cách. Đầu tiên, chúng tôi nhắm mục tiêu vào một nhiệm vụ khác, tức là, chúng tôi tập trung vào đề xuất trang phục chứ không phải xếp hạng điểm. Thứ hai, khuyến nghị và thế hệ trong bài viết này không được cá nhân hóa. Chúng tôi xác định xem trang phục có phù hợp hay không dựa trên quan điểm chung. Bởi vì các yếu tố ảnh hưởng đến mọi người Lựa chọn quần áo của chủ yếu bao gồm thời trang, nghề nghiệp, tuổi tác và chủ nghĩa khu vực hiện tại và chúng tôi tin rằng những người có độ tuổi và mức độ phổ biến tương tự nhau thường giống nhau về các yếu tố này. Hơn nữa, các bình luận được tạo ra từ nhiều bình luận (từ những người dùng trực tuyến khác nhau) phản ánh ý kiến ​​chung và chung thay mặt cho nhiều người dùng thay vì một người dùng cụ thể. Thứ ba, không giống như Ni et al. [20] Công việc của ho và Li et al. [21] Công việc của séc, nhiệm vụ của chúng tôi liên quan đến nhiều phương thức (nghĩa là hình ảnh và văn bản). Thứ tư, thay vì sử dụng RNN đơn giản, chúng tôi đề xuất một cơ chế chú ý đa phương thức phức tạp hơn để xử lý ánh xạ từ trực quan đến không gian văn bản.

* 1. **Tạo văn bản**

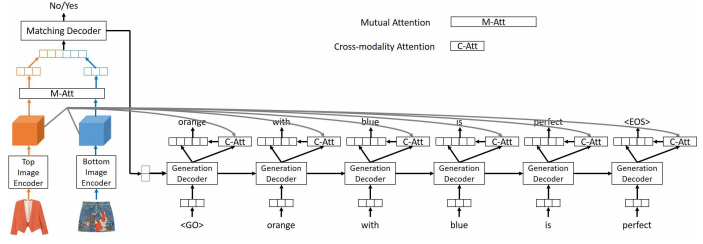
Tạo văn bản bao gồm nhiều nhiệm vụ và nghiên cứu khác nhau, chẳng hạn như tóm tắt văn bản [22, 23], dịch máy [24, 25], hệ thống đối thoại [26, 27] và chú thích hình ảnh [28, 29]. Chúng tôi liệt kê một số công việc liên quan về bình luận hoặc đánh giá thế hệ như sau.

Cao et al. [30] trình bày một khung để tự động thu thập các blog nhỏ có liên quan từ các trang web blog để tạo bình luận cho các tin tức phổ biến trên các trang web tin tức. Lipton và cộng sự. [31] thiết kế RNN cấp độ ký tự để tạo đánh giá. Các đánh giá được tạo dựa trên thông tin phụ trợ, chẳng hạn như ID người dùng / mặt hàng, danh mục và xếp hạng. Radford và cộng sự. [32] cũng huấn luyện một mô hình ngôn ngữ RNN cấp độ ký tự trên bộ dữ liệu đánh giá Amazon, chỉ có một lớp LSTM nhân với một đơn vị ẩn với 4.096 đơn vị ẩn. Họ giới thiệu một đơn vị đặc biệt trong số các đơn vị ẩn có thể kiểm soát tình cảm của các đánh giá được tạo. Đồng và cộng sự. [33] đề xuất một mô hình thuộc tính theo trình tự để tạo ra các đánh giá sản phẩm cho thông tin thuộc tính đã cho bao gồm người dùng, sản phẩm và xếp hạng. Đầu tiên họ sử dụng một bộ mã hóa thuộc tính để tìm hiểu các vectơ biểu diễn của các thuộc tính đầu vào. Sau đó, họ sử dụng một LSTM xếp chồng với cơ chế chú ý để tạo ra các đánh giá dựa trên các vectơ đại diện này. Tang et al. [34] đề xuất hai cách tiếp cận mới, lần đầu tiên mã hóa bối cảnh, chẳng hạn như xếp hạng tình cảm và id sản phẩm, thành một biểu diễn ngữ nghĩa liên tục và sau đó giải mã biểu diễn ngữ nghĩa thành các đánh giá với RNNs. Hu et al. [35] kết hợp một bộ mã hóa tự động đa dạng và một bộ phân biệt thuộc tính tổng thể để tạo ra các đánh giá. Họ luân phiên huấn luyện bộ mã hóa tự động và bộ phân biệt đối xử. Họ có thể tự động kiểm soát các thuộc tính của các đánh giá được tạo bằng cách tìm hiểu các biểu diễn tiềm ẩn không đồng nhất với ngữ nghĩa được chỉ định. Những nghiên cứu này chỉ tập trung vào việc tạo văn bản, và không cùng thực hiện khuyến nghị.

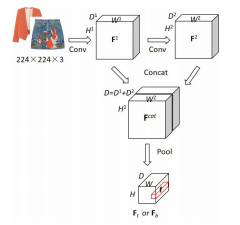
1. **NEURAL OUTFIT RECOMMENDATION**
   1. **Tổng quan**

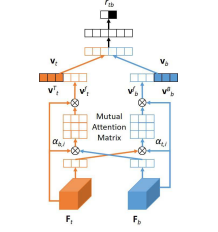
Cho một phần trên *ti* từ một nhóm T = {t1, t2 ,. . . , tNt}, nhiệm vụ đề xuất phần dưới là đề xuất một danh sách các phần dưới được xếp hạng từ nhóm ứng cử viên B = {b1, b2 ,. . . , bNb}. Tương tự, nhiệm vụ đề xuất phần trên là đề xuất một danh sách các phần trên được xếp hạng cho một phần dưới nhất định. Nhiệm vụ tạo bình luận là tạo ra một bình luận nghe có vẻ tự nhiên *ctb* cho mỗi trang phục được đề xuất (nghĩa là, cặp phía trên). Các ý kiến được tạo ra có thể được coi là lời giải thích cho từng trang phục được đề xuất: tại sao một trang phục phù hợp? Lưu ý rằng việc chúng tôi thực hiện đề xuất phần dưới hay đề xuất phần trên không quan trọng, NOR tạo ra các nhận xét tương tự cho cùng một trang phục, bởi vì các nhận xét được tạo là cho toàn bộ trang phục.

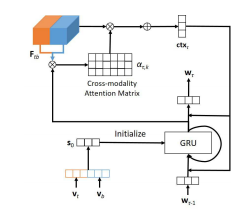
Như trong Hình 2, NOR bao gồm ba thành phần cốt lõi, bộ mã hóa hình ảnh trên và dưới, bộ giải mã phù hợp và bộ giải mã thế hệ. Dựa trên mạng nơ ron tích chập [36], bộ mã hóa hình ảnh trên và dưới (Hình 3 (a)) trích xuất các đặc điểm hình ảnh từ các hình ảnh bao gồm một cặp (t, b) và chuyển đổi các đặc điểm hình ảnh thành các biểu diễn tiềm ẩn của t và b , tương ứng. Một cơ chế chú ý lẫn nhau được giới thiệu ở đây để đảm bảo rằng bộ mã hóa hình ảnh trên và dưới có thể mã hóa tính tương thích giữa t và b thành các biểu diễn tiềm ẩn của chúng. Trong hình 3 (b), bộ giải mã phù hợp là một tri giác đa lớp (MLP) để đánh giá điểm phù hợp giữa t và b. Bộ giải mã thế hệ trong Hình 3 (c) là một đơn vị lặp lại có kiểm soát (GRU) [37], được sử dụng để dịch sự kết hợp giữa biểu diễn tiềm ẩn của đỉnh và biểu diễn tiềm ẩn của đáy thành một chuỗi các từ dưới dạng nhận xét . Đối với bộ giải mã thế hệ, chúng tôi đề xuất sự chú ý đa phương thức để mô hình hóa tốt hơn sự chuyển đổi giữa không gian hình ảnh và văn bản.



Hình 2: Tổng quan về kiến trúc đề xuất trang phục thần kinh (NOR) được đề xuất. NOR chứa ba phần: (1) bộ mã hóa hình ảnh trên và dưới (tương ứng với Hình 3 (a)), (2) bộ giải mã phù hợp (tương ứng với Hình 3 (b)) và (3) bộ giải mã thế hệ ( tương ứng với hình 3 (c)).

(a) Bộ mã hóa hình ảnh trên và dưới.

(b) Sự chú ý lẫn nhau và bộ giải mã phù hợp.

(c) Sự chú ý đa phương thức và bộ giải mã thế hệ.

Hình 3: Chi tiết về kiến trúc đề xuất thời trang thần kinh (NOR). (a) Bộ mã hóa hình ảnh trên và dưới trích xuất các tính năng hình ảnh *Ft* và *Fb* từ hình ảnh. (b) Sử dụng cơ chế chú ý lẫn nhau, chúng tôi chuyển đổi các tính năng trực quan sang các biểu diễn tiềm ẩn *vt* và *vb*. Sau đó, bộ giải mã phù hợp dự đoán chỉ số phù hợp *rtb*. (c) Tại mỗi mốc thời gian, bộ giải mã thế hệ sử dụng cơ chế chú ý đa phương thức để tạo ra từ *wτ*.

* 1. **Bộ mã hóa hình ảnh trên và dưới**

Bộ mã hóa hình ảnh trên cùng và bộ mã hóa hình ảnh phía dưới là CNN, được sử dụng rộng rãi trong khuyến nghị trang phục [8, 13]. Mặc dù có nhiều kiến trúc mạnh mẽ, như ResNet [38] hoặc DenseNet [39], việc đào tạo các kiến trúc này không dễ dàng, bởi vì chúng có nhiều tham số và cần nhiều dữ liệu và thời gian để đào tạo. Để cân bằng chi phí đào tạo và hiệu suất thử nghiệm, chúng tôi thiết kế một CNN hai lớp với sự chú ý lẫn nhau là bộ mã hóa hình ảnh trên cùng và dưới cùng, có ít tham số hơn và mang lại hiệu suất tốt. Chúng tôi xác minh tính hiệu quả của kiến trúc mạng của chúng tôi trong các thử nghiệm.

Cho một cặp hình ảnh (It, Ib), chúng tôi giả sử rằng hình ảnh It và hình ảnh Ib có kích thước 224 × 224 với 3 kênh. Như được hiển thị trong Hình 3 (a), chúng tôi trích xuất các tính năng trực quan từ It hoặc Ib thông qua CNN hai lớp. Cụ thể, trước tiên chúng tôi cung cấp It hoặc Ib cho một lớp chập để có các tính năng trực quan chính F1 ∈ R H1 × W1 × D1. Sau đó, chúng tôi đưa F1 vào một lớp chập khác để có được các tính năng hình ảnh tiên tiến F2 ∈ R H2 × W2 × D2. Đưa ra các bài học với DensetNet [39] để sử dụng các tính năng hình ảnh hiệu quả hơn trong các lớp CNN khác nhau, chúng tôi đảm bảo rằng H1 = H2 và W1 = W2 với các thao tác đệm để chúng ta có thể ghép nối F1 và F2 để có được Fcat ∈ R H2 × W2 × (D1 + D2). Cuối cùng, chúng tôi sử dụng nhóm tối đa trong Fcat để có được các tính năng hình ảnh cuối cùng F ∈ RH × W × D.

Sau đó, chúng tôi định hình lại F = [f 1 ,. . . , f L] bằng cách làm phẳng chiều rộng và chiều cao của F ban đầu, trong đó f i ∈ R D và L = W × H. Chúng ta có thể coi f i là các tính năng trực quan của khu vực địa phương thứ i của hình ảnh đầu vào. Cho một cặp bao gồm hình ảnh trên cùng I và hình dưới cùng Ib, chúng sẽ được đưa vào cùng một CNN, tức là, bộ mã hóa hình ảnh trên cùng và dưới cùng có cùng cấu trúc và các tham số chia sẻ. Đối với nó, các tính năng hình ảnh được trích xuất Ft được ký hiệu như trong biểu thức. 1:



Tương tự, đối với các tính năng hình ảnh được trích xuất Fb của hình ảnh Ib, chúng ta có:

