ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

VƯƠNG THỊ HỒNG

TRUY HỒI CHÉO MÔ HÌNH CHO NHẠC VÀ LỜI BÀI HÁT

Ngành: Hệ thống thông tin

Chuyên ngành: Hệ thống thông tin

Mã Số: 8480104.01

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC:

PGS.TS. HÀ QUANG THỤY

Hà nội – 12/2018

Mục lục

LỜI CẢM ƠN

Trước tiên tôi xin dành lời cảm ơn chân thành và sâu sắc đến thầy giáo PGS. TS. Hà Quang Thụy – người đã hướng dẫn, khuyến khích, chỉ bảo và tạo cho tôi những điều kiện tốt nhất từ khi bắt đầu cho tới khi hoàn thành công việc của mình.

Tôi cũng xin chân thành cảm ơn TS. Yi Yu – giảng viên Viện tin học quốc gia, Nhật Bản đã tạo điều kiện tốt nhất cho tôi hoàn thành chương trình thực tập cao học. Đồng thời tôi xin chân thành cảm ơn thầy cô và anh chị Phòng thí nghiệm Công nghệ và tri thức đã giúp đỡ, động viên tôi trong thời gian học tập và công tác.

Tôi xin dành lời cảm ơn chân thành tới các thầy cô giáo khoa Công nghệ thông tin, trường Đại học Công nghệ, ĐHQGHN đã tận tình đào tạo, cung cấp cho tôi những kiến thức vô cùng quý giá và đã tạo điều kiện tốt nhất cho tôi trong suốt quá trình học tập, nghiên cứu tại trường.

Cuối cùng, tôi xin cảm ơn tất cả những người thân yêu trong gia đình tôi cùng toàn thể bạn bè những người đã luôn giúp đỡ, động viên tôi học tập và nghiên cứu chương trình thạc sĩ tại Đại học Công nghệ, ĐHQGHN.

LỜI CAM ĐOAN

Tôi xin cam đoan rằng luận văn thạc sĩ công nghệ thông tin "Truy hồi chéo mô hình cho nhạc và lời bài hát" là công trình nghiên cứu của riêng tôi, không sao chép lại của người khác. Trong toàn bộ nội dung của luận văn, những điều đã được trình bày hoặc là của chính cá nhân tôi hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các nguồn tài liệu tham khảo đều có xuất xứ rõ ràng và hợp pháp.

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

Hà Nội, ngày …. tháng 12 năm 2018

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

LỜI MỞ ĐẦU

Ngày nay, dữ liệu đa phương tiện phát triển nhanh chóng bởi các trang mạng ngày càng cập nhật nhiều tin tức mang tính thời sự cũng như mang tính sở thích cá nhân hóa với sự đa dạng các kiểu dữ liệu văn bản, hình ảnh hay âm thanh. Các kiểu dữ liệu như văn bản, hình ảnh và âm thanh được sử dụng cùng nhau đều mô tả cùng sự kiện hoặc cùng chủ đề được đề cập tới gọi là dữ liệu đa phương thức [16]. Dữ liệu đa phương thức được ứng dụng cho truy hồi chéo mô hình, hệ tư vấn hoặc phát hiện chủ đề ẩn. Những năm gần đây, truy hồi chéo mô hình đã trở thành xu hướng nghiên cứu của cộng đồng. Nhiều nghiên cứu trên thế giới như [3, 5, 14, 18] tập trung vào truy hồi chéo mô hình cho văn bản và hình ảnh, video và hình ảnh. Các phương pháp truy hồi cổ điển chỉ dựa vào một mô hình [2, 7, 11], những kỹ thuật chỉ sử dụng siêu dữ liệu (meta data) như từ khóa, thẻ hoặc đoạn mô tả nội dung liên quan hơn là dựa vào chính nội dung của dữ liệu đa phương thức. Các nghiên cứu [18, 20, 21] tập trung đề xuất các ý tưởng sử dụng học sâu để truy hồi chéo mô hình tăng hiệu quả về độ chính xác dựa trên chính nội dung của dữ liệu đa phương thức.

Truy hồi chéo mô hình không chỉ là chủ đề quan tâm của cộng đồng nghiên cứu thế giới mà còn nhận sự quan tâm của công nghiệp. Các nghiên cứu và ứng dụng nhằm cải tiến và đáp ứng được nhu cầu truy vấn chéo thông tin giữa các dữ liệu đa phương thức của người dùng. Cùng góp phần vào trào lưu nghiên cứu thế giới, luận văn có tên đề tài truy hồi chéo mô hình cho nhạc và lời bài hát thực hiện để xây dựng mô hình cho phép truy hồi chéo khi sử dụng nhạc là truy vấn hoặc khi sử dụng lời bài hát là truy vấn. Xuất phát từ ứng dụng thực tế cần xây dựng hệ thống truy hồi chéo thông tin của các dữ liệu đa phương tiện cho phép truy vấn chéo giữa các kiểu dữ liệu khác nhau. Luận văn tập trung vào giải quyết bài toán cải tiến độ chính xác truy hồi chéo mô hình cho nhạc và lời bài hát. Phương pháp tiếp cận trong luận văn kết hợp học sâu và phân tích tương quan chính tắc để cải tiến độ chính xác cho mô hình.

Luận văn gồm bốn chương nội dung được mô tả sơ bộ như sau:

Chương 1. *Giới thiệu truy hồi thông tin* trình bày dữ liệu đa phương thức, truy hồi thông tin nói chung và truy hồi chéo mô hình nói riêng. Trình bày sơ lược phân loại truy hồi chéo mô hình và quy trình chung để giải quyết bài toán truy hồi chéo mô hình, đồng thời phát biểu bài toán của luận văn triển khai.

Chương 2. *Các phương pháp truy hồi chéo mô hình* trình bày hai phương pháp chính: phương pháp học không gian con, học sâu và một số phương pháp khác cho truy hồi chéo mô hình.

Chương 3. *Mô hình đề xuất* trình bày phương pháp tiếp cận bài toán và đưa ra quy trình xây dựng mô hình và các chi tiết từng pha. Chương này chỉ ra cách thực hiện các bước trong mô hình dựa trên cách tiếp cận của phương pháp đã đề xuất.

Chương 4. *Thực nghiệm và đánh giá* mô tả dữ liệu, trích xuất đặc trưng cho từng kiểu dữ liệu, môi trường và công cụ thực nghiệm. Đồng thời chương 4 mô tả kịch bản thực nghiệm, đưa ra kết quả và đánh giá mô hình đề xuất.

Cuối cùng, phần kết luận đưa ra nhận xét và đánh giá chung về kết quả đạt được của luận văn.

Chương 1: Giới thiệu truy hồi thông tin

Chương 1 tập trung vào giới thiệu về dữ liệu đa phương thức, truy hồi thông tin nói chung và truy hồi chéo mô hình nói riêng. Trình bày sơ lược phân loại truy hồi chéo mô hình và quy trình chung để giải quyết bài toán truy hồi chéo mô hình, đồng thời phát biểu bài toán của luận văn triển khai.

1.1 Dữ liệu đa phương thức và truy hồi thông tin

Hơn thập kỉ qua, dữ liệu đa phương tiện phát triển nhanh chóng và gia tăng bởi số lượng người dùng ngày càng lớn. Các trang mạng ngày càng cập nhật nhiều tin tức vừa mang tính thời sự vừa mang tính sở thích cá nhân hóa với sự đa dạng các kiểu dữ liệu văn bản, hình ảnh hay âm thanh. Đối với các trang mạng xã hội, dữ liệu được tạo ra bởi cộng đồng người dùng, người dùng có thể tự đăng bài có nội dung là văn bản, hình ảnh hoặc video mà không giới hạn về số lượng nội dung hoặc bài đăng trong ngày. Các kiểu dữ liệu như văn bản, hình ảnh và âm thanh được sử dụng cùng nhau đều mô tả cùng sự kiện hoặc cùng chủ đề được đề cập tới gọi là *dữ liệu đa phương thức* (multi-modal data) [16]. Sự phát triển nhanh chóng của mạng xã hội cho phép cộng đồng kết nối, chia sẻ và giao tiếp với nhau một cách dễ dàng. Theo thống kê của Facebook[1] đến hết tháng 9 năm 2014 số lượng người dùng hoạt động là 890 triệu người, tăng 18% so với cùng kì năm 2013. Đến nay, con số thống kê người dùng Facebook lên hơn 1 tỉ người dùng trên toàn thế giới. Instagram là ứng dụng cộng đồng cho phép đăng văn bản ngắn và hình ảnh thu hút hơn 1 tỉ người dùng tính tới tháng 6 năm 2018. Chính vì sự gia tăng dữ liệu đa phương thức nói chung và dữ liệu đa phương tiện nói riêng, người dùng sẽ gặp khó khăn trong việc tìm kiếm thông tin liên quan một cách hiệu quả và nhanh chóng.

Dữ liệu đa phương thức được ứng dụng cho truy hồi chéo mô hình, hệ tư vấn hoặc phát hiện chủ đề ẩn [16]. Dữ liệu dạng hình ảnh, âm thanh hay văn bản cùng đề cập tới một sự kiện, chủ đề thì giữa chúng có mối tương quan ngữ nghĩa. Ứng dụng dữ liệu đa phương thức cho truy hồi chéo mô hình giữa ảnh và văn bản [17, 21] , cho âm nhạc giữa nhạc và lời bài hát [20]. Bên cạnh sự phát triển của dữ liệu đa phương thức, các phương pháp, kỹ thuật để lập

[1] http://investor.fb.com/annuals.cfm

chỉ mục và tìm kiếm dữ liệu đa phương thức được quan tâm nghiên cứu. Tuy nhiên, các kỹ thuật tìm kiếm này chủ yếu dựa trên mô hình dựa trên từ khóa hoặc nội dung truy xuất cho phép thực hiện tìm kiếm tương tự trên cùng một loại dữ liệu, ví dụ truy hồi văn bản, truy hồi hình ảnh, truy hồi [2, 7, 11]. Do đó, một yêu cầu đòi hỏi để thúc đẩy truy hồi thông tin là phát triển một mô hình truy hồi mới có thể hỗ trợ tìm kiếm tương tự cho nhiều kiểu dữ liệu đề cập tới cùng chủ đề hay sự kiện.

Những năm gần đây, truy hồi chéo mô hình hay truy hồi chéo thông tin đã trở thành xu hướng nghiên cứu bởi sự phát triển nhanh chóng của dữ liệu đa phương thức. Truy hồi chéo mô hình sử dụng một kiểu dữ liệu như truy vấn để truy xuất những kiểu dữ liệu khác liên quan. Ví dụ, một người dùng có thể sử dụng một đoạn văn bản ngắn truy vấn để tìm ra danh sách các hình ảnh hoặc âm thanh phù hợp với đoạn văn bản ngắn và ngược lại, sử dụng một hình ảnh hoặc âm thanh truy vấn để tìm ra những danh sách các từ liên quan nhất tới hình ảnh hoặc âm thanh. Các ứng dụng mạng xã hội như Facebook, Flickr, Youtube và Twitter đang thay đổi cách mọi người tương tác với thế giới và thông tin quan tâm. Người dùng gửi nội dung bất kì của một kiểu dữ liệu nào đó để truy vấn một kiểu dữ liệu khác sao cho đều có cùng ngữ nghĩa. Do đó, việc truy hồi chéo mô hình ngày càng trở nên quan trọng. Nhiều nghiên cứu trên thế giới, như [3, 5, 14, 18] tập trung vào truy hồi mô hình chéo cho văn bản và hình ảnh, video và hình ảnh. Thách thức của truy hồi chéo mô hình là làm sao để đo được sự tương tự nội dung giữa các kiểu dữ liệu khác nhau. Các phương pháp truy hồi cổ điển chỉ dựa vào một mô hình [2, 7, 11], những kỹ thuật chỉ sử dụng siêu dữ liệu (meta data) như từ khóa, thẻ hoặc đoạn mô tả nội dung liên quan hơn là dựa vào chính nội dung của dữ liệu đa phương thức. Các phương pháp truy hồi chéo mô hình yêu cầu phải mô hình hóa mối quan hệ giữa các kiểu dữ liệu để người dùng có thể tìm được những gì liên quan nhất tới truy vấn của họ. Các nghiên cứu [18, 20, 21] tập trung đề xuất các ý tưởng truy hồi chéo mô hình tăng hiệu quả về độ chính xác dựa trên chính nội dung của dữ liệu đa phương thức.

1.2 Phân loại truy hồi chéo mô hình

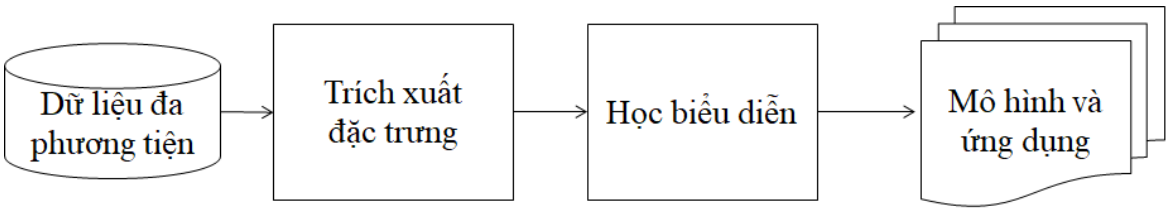
Đối với truy hồi chéo mô hình dựa trên nội dung của dữ liệu đa phương thức, theo nhóm tác giả Wang và cộng sự [16], truy hồi thông tin chéo được chia thành 2 loại chính dựa trên học biểu diễn là giá trị thực [13, 14, 18] và học biểu diễn là giá trị nhị phân [5, 17, 22]. Truy hồi thông tin chéo dựa trên biểu diễn giá trị thực, không gian biểu diễn chung được học cho các kiểu dữ liệu là giá trị thực được trích xuất dựa trên chính nội dung của kiểu dữ liệu đó. Còn với truy hồi thông tin chéo dựa trên biểu diễn giá trị nhị phân, không gian biểu diễn chung được học cho các kiểu dữ liệu là giá trị nhị phân với bit 0 và bit 1 được chuyển đổi từ nội dung dữ liệu tương ứng. Phương pháp biểu diễn học nhị phân mục tiêu chuyển đổi các kiểu dữ liệu khác nhau thành một không gian Hamming chung. Do đó, các ứng dụng thực tiễn mà quan trọng tốc độ xử lý sẽ ưu tiên việc sử dụng phương pháp học biểu diễn nhị phân. Tuy nhiên, với việc biểu diễn là mã hóa các mã nhị phân nên độ chính xác truy hồi thường giảm nhẹ do mất mát thông tin trong quá trình mã hóa. Tuy nhiên với các ứng dụng thực tiễn mà quan trọng độ chính xác của truy hồi thông tin được ưu tiên hơn nên sử dụng phương pháp học biểu diễn giá trị thực. Tùy vào mục đích thực tiễn ưu tiên tốc độ hay độ chính xác mà lựa chọn phương pháp học biểu diễn dựa trên giá trị thực hay nhị phân. Khóa luận tập trung vào truy hồi chéo mô hình dựa trên học giá trị thực bằng cách trích xuất đặc trưng của dữ liệu đa phương thức bằng các kỹ thuật học máy dựa trên chính nội dung của dữ liệu.

Dựa trên học biểu diễn để tìm ra không gian chung cho phép ánh xạ giữa các dữ liệu đa phương thức, các phương pháp truy hồi chéo mô hình theo [16] có thể được chia thành bốn nhóm: không giám sát (unsupervised), giám sát (supervised), phương pháp dựa trên từng cặp (pairwise method), phương pháp dựa trên xếp hạng (rank based method). Phương pháp học không giám sát chỉ có các thông tin của dữ liệu đa phương thức dùng để học biểu diễn chung mà không có nhãn. Còn phương pháp học giám sát sử dụng nhãn và các thông tin của dữ liệu liệu đa phương thức để học biểu diễn chung. Như vậy vai trò của nhãn cũng góp phần xây dựng không gian học biểu diễn ý nghĩa về nhãn thay vì chỉ sử dụng nội dung từ chính dữ liệu đa phương thức. Phương pháp học

dựa trên từng cặp đầu vào sử dụng các cặp tương tự nhau hoặc các cặp phân biệt nhau của chính dữ liệu đa phương thứ để thực hiện học đại diện chung. Những phương pháp này học khoảng cách số liệu có ý nghĩa giữa các dữ liệu của mô hình khác nhau. Còn phương pháp học dựa trên xếp hạng, danh sách xếp hạng được thực hiện để học đại diện chung. Các phương pháp học xếp hạng sẽ quan tâm tới độ đo để tính toán hạng của dữ liệu đa phương thức. Phương pháp học dựa trên xếp hạng cũng được nghiên cứu cho truy hồi chéo mô hình ở pha thứ ba như một bài toán của học xếp hạng. Các kỹ thuật điển hình cho truy hồi chéo mô hình như phân tích tương quan chính tắc (canonical correlation analysis/ CCA), học tương quan chính tắc sâu (deep canonical correlation analysis/DCCA), mô hình chủ đề ẩn.

Trong hệ thống truy hồi chéo mô hình, người dùng có thể tìm kiếm chéo giữa dữ liệu đa phương thức, ví dụ sử dụng văn bản như truy vấn để truy xuất hình ảnh hay sử dụng hình ảnh như truy vấn để truy xuất văn bản liên quan [13, 14, 18] hoặc sử dụng hình ảnh như truy vấn để truy xuất video và ngược lại. Nếu các kiểu dữ liệu liên quan đến cùng sự kiện hoặc chủ đề, chúng được kì vọng là chia sẻ không gian đại diện chung – nơi mà có thể đo trực tiếp được sự tương tự giữa các dữ liệu đa phương thức. Theo [16], kiến trúc chung của hệ thống truy hồi mô hình chéo gồm ba pha được minh họa trong hình 1.1: trích xuất đặc trưng, học biểu diễn (representation learning), mô hình và ứng dụng. Pha thứ nhất trích xuất đặc trưng là lựa chọn đặc trưng biểu diễn cho từng kiểu dữ liệu. Tùy thuộc là kiểu dữ liệu văn bản, hình ảnh hay âm thanh thì sẽ có các kỹ thuật xử lý trích chọn đặc trưng và lựa chọn sử dụng đặc trưng nào cho bài toán. Ví dụ đối với văn bản, đặc trưng túi từ (Bag of Word) thường được sử dụng, hình ảnh thường sử dụng điểm ảnh nhị phân để biểu diễn đặc trưng, âm thanh thì sử dụng đặc trưng phổ để biểu diễn. Pha thứ hai là học biểu diễn dữ liệu, mô hình hóa chéo sự tương tự được thực hiện để học ra đại diện cho các kiểu dữ liệu khác nhau theo bốn phương pháp tiếp cận là học giám sát, không giám sát, theo cặp và xếp hạng. Trong không gian biểu diễn, kiểu dữ liệu này sẽ được sử dụng như truy vấn để truy xuất tới kiểu dữ liệu khác. Pha cuối cùng là ứng dụng, sử dụng học biểu diễn cho phép truy hồi mô hình chéo bằng cách xếp hạng kết quả tìm kiếm trả về. Vì các đặc trưng của các kiểu dữ liệu khác nhau thường có sự phân phối và biểu diễn

không nhất quán nên cần phải có cầu nối – nơi mà có thể tìm được sự tương tự về mặt ngữ nghĩa của chéo mô hình. Một cách tiếp cận phổ biến nhất là học biểu diễn, mục tiêu là tìm các ánh xạ đặc trưng của các mô hình khác nhau trong không gian đại diện đặc trưng chung.



Hình 1.1: Quy trình truy hồi chéo mô hình cho dữ liệu đa phương tiện

1.3 Phát biểu bài toán

Để tận dụng tối đa dữ liệu đa phương tiện nói chung và sử dụng tối ưu công nghệ đa phương tiện đang phát triển nhanh chóng, các cơ chế tự động là cần thiết để thiết lập một liên kết tương tự từ một dữ liệu dạng này sang một dữ liệu dạng khác nếu chúng có liên quan ngữ nghĩa. Xuất phát từ ứng dụng thực tế cần xây dựng hệ thống truy hồi chéo thông tin của các dữ liệu đa phương tiện cho phép truy vấn chéo giữa các kiểu dữ liệu khác nhau. Có nhiều kiểu dữ liệu khác nhau như văn bản, hình ảnh, âm thanh được ứng dụng cho truy hồi chéo. Mỗi kiểu dữ liệu khác nhau, đòi hỏi kỹ thuật trích chọn đặc trưng khác nhau. Luận văn tập trung vào giải quyết bài toán cải tiến độ chính xác truy hồi chéo mô hình cho nhạc và lời bài hát.

Ý nghĩa: Truy hồi chéo mô hình không chỉ là chủ đề quan tâm của cộng đồng nghiên cứu thế giới mà còn nhận sự quan tâm của công nghiệp. Các nghiên cứu và ứng dụng nhằm cải tiến và đáp ứng được nhu cầu truy vấn chéo thông tin giữa các dữ liệu đa phương thức của người dùng. Cùng góp phần vào trào lưu nghiên cứu thế giới, luận văn có tên đề tài truy hồi chéo mô hình cho nhạc và lời bài hát thực hiện để xây dựng mô hình cho phép truy hồi chéo khi sử dụng nhạc là truy vấn hoặc khi sử dụng lời bài hát là truy vấn. Mô hình cho phép sử dụng nhạc như truy vấn và truy xuất ra danh sách các lời bài hát đã được xếp hạng và ngược lại, sử dụng lời bài hát như truy vấn và truy xuất ra danh sách các nhạc đã được xếp hạng. Ứng dụng mô hình đề xuất trong luận văn có thể xây dựng các trang web tìm kiếm âm nhạc hiệu quả cho người dùng hoặc nhúng mô hình vào hệ thống các trang web âm nhạc có sẵn.

Đầu vào: Tập các dữ liệu nhạc, dữ liệu lời bài hát và nhãn cảm xúc tương ứng với mỗi cặp dữ liệu.

Đầu ra: Mô hình học biểu diễn cho nhạc và lời bài hát. Sử dụng mô hình này để truy hồi chéo mô hình cho nhạc và lời bài hát. Cụ thể luận văn giải quyết hai bài toán con:

1. Xây dựng mô hình cho phép truy hồi thông tin chéo giữa nhạc và lời bài hát. Cụ thể tìm ra được không gian biểu diễn $\boldsymbol{S} = \{\boldsymbol{S}_A, \boldsymbol{S}_T\}$ với 2 hàm không gian biểu diễn với *d* chiều cho nhạc và lời bài hát được ánh xạ bởi hàm f_A, f_T : $\boldsymbol{S}_A = f_A(\mathbf{A}, \theta_A)$, $\boldsymbol{S}_T = f_T(\mathbf{T}, \theta_T)$, trong đó θ_A , θ_T là các tham số học cho nhạc, lời bài hát tương ứng.
2. Sử dụng mô hình biểu diễn cho truy hồi chéo mô hình và đánh giá hiệu quả mô hình bằng độ đo xếp hạng.

Một số kí hiệu, khái niệm được sử dụng trong luận văn được giải thích trong bảng 1.1.

Bảng 1.1: Bảng các kí hiệu và giải thích

STT	Ký hiệu	Giải thích
1	$\mathbf{I} = \{ I_1, I_2, \ldots, I_n\}$ vs $I_i = (a_i, t_i)$	Tập *n* cặp, mỗi cặp là nhạc và lời bài hát tương ứng
2	$\mathbf{A} = \{ a_1, a_2, \ldots, a_n\}$, $a_i \in \mathbb{R}^{d1}$	Tập *n* vector audio với *d1* chiều
3	$\mathbf{T} = \{ t_1, t_2, \ldots, t_n\}$, $t_i \in \mathbb{R}^{d2}$	Tập *n* vector lời nhạc với *d2* chiều
4	$\mathbf{Y} = \{ y_1, y_2, \ldots, y_n\}$, $y_i = \{ y_{i1}, y_{i2}, \ldots, y_{ic}\} \in \mathbb{R}^c$, c =20	Tập *n* vector lời nhạc với *c* chiều Nhãn cảm xúc của mỗi cặp nhạc và lời bài hát
5	$\boldsymbol{S} = \{\boldsymbol{S}_A, \boldsymbol{S}_T\}$ $\boldsymbol{S}_A = f_A(\mathbf{A}, \theta_A)$, $\boldsymbol{S}_T = f_T(\mathbf{T}, \theta_T)$,	Không gian biểu diễn với *d* chiều cho nhạc và lời bài hát được ánh xạ bởi hàm f_A, f_T

Luận văn nhằm mục đích nghiên cứu phương pháp xây dựng hệ thống truy hồi chéo mô hình cho nhạc và lời bài hát. Bên cạnh đó, luận văn cũng đề xuất phương pháp mới để cải tiến hiệu quả độ chính xác của hệ thống truy hồi chéo mô hình. Phương pháp đề xuất luận văn có thể được mở rộng áp dụng cho các miền dữ liệu phương thức khác như cho ảnh và văn bản, ảnh và video trong bài toán truy hồi chéo.

Chương 2: Các phương pháp truy hồi chéo mô hình

Dữ liệu của các mô hình khác nhau liên quan đến cùng sự kiện, chủ đề thì giữa chúng được dự đoán là cùng chia sẻ không gian đại diện chung – nơi mà dữ liệu liên quan là gần nhau trong không gian. Các phương pháp học biểu diễn dựa trên giá trị thực hay giá trị nhị phân đều có mục đích học một không gian biểu diễn chung nội dung – nơi mà dữ liệu các mô hình khác nhau có thể so sánh trực tiếp. Dựa theo việc cung cấp thông tin đầu vào cho việc học, phương pháp học biểu diễn được chia bốn loại: học giám sát, học bám giám sát, học từng cặp, học xếp hạng như đã trình bày Chương 1. Chương 2 trình bày các kỹ thuật điển hình cho các phương pháp học biểu diễn trên.

2. 1 Phương pháp học không gian con

Tính toán đo được sự tương tự giữa các dữ liệu mô hình khác nhau cho truy hồi chéo mô hình là bài toán khó. Phương pháp học không gian con là một phương pháp phổ biến nhất. Mục đích của phương pháp này là tìm được không gian chung chia sẻ bởi dữ liệu các mô hình khác nhau. Học không gian con bán giám sát sử dụng thông tin cặp để học ra không gian ẩn chung cho dữ liệu đa phương thức. Chúng buộc các cặp gần nhau giữa các dữ liệu đa phương thức thành không gian chung. Phân tích tương quan chính tắc (Canonical Correlation Analyis) CCA là một phương pháp học không gian để xác định mối quan hệ chéo mô hình giữa các dữ liệu từ các mô hình khác nhau. CCA là một phương pháp thống kê thăm dò phổ biến, cho phép phân tích các mối quan hệ tồn tại giữa hai tập biến. Việc chuyển đổi tuyến tính tốt nhất cho hai tập dữ liệu đa chiều, cho phép tương quan tối đa giữa chúng có thể đạt được bằng sử dụng CCA. CCA đã được áp dụng thành công cho nhiều lĩnh vực khoa học y sinh quan trọng cũng như được sử dụng rộng rãi cho bài toán truy hồi chéo đa phương thức [18, 19, 20]. CCA học tương quan giữa hai dữ liệu (x,y) đa phương thức là lớn nhất theo công thức (2.1) :

$$\mathbf{max}_{\mathbf{w_x w_y}} = \frac{\mathbf{w_x^T C_{xy} w_y}}{\sqrt{\mathbf{w_x^T C_{xx} w_x}}\sqrt{\mathbf{w_y^T C_{yy} w_y}}} \qquad (2.1)$$

trong đó, C_{xx}, C_{yy}, C_{xy} là ma trận hiệp phương sai của dữ liệu x, dữ liệu y, dữ liệu x và y tương ứng. CCA học không gian ngữ nghĩa chung để tính toán độ tương tự của các đặc trưng mô hình khác nhau.

Cho p và q là số lượng các đặc trưng của hai tập dữ liệu đa biến X và Y tương ứng, trong đó số lượng mẫu trong cả X và Y là *n*. Công nghệ hiện đại cho phép nhiều hướng hơn trên luồng dữ liệu, điều này xảy ra trong không gian đặc trưng chiều rất cao p và q. Mặt khác, số lượng mẫu đào tạo *n* thường bị giới hạn. Khi $n << (p, q)$ các đặc trưng trong X và Y có khuynh hướng được đánh giá cao, điều này dẫn đến điều kiện không tốt của ma trận hiệp phương sai C_{xx}, C_{yy} của X và Y tương ứng [9]. Thực tế nghịch đảo của chúng không còn đáng tin cậy nữa, dẫn đến việc tính toán CCA không có giá trị. Có hai cách để khác phục vấn đề này. Cách tiếp cận đầu tiên là phiên bản CCA chuẩn hóa. Theo [9], trong CCA chuẩn hóa (regularized CCA:RCCA), các phần tử đường chéo của ma trận hiệp phương sai C_{xx}, C_{yy} phải được tăng lên bằng cách sử dụng tối ưu hóa tìm kiếm lưới. Mặc khác, các phần tử ngoài đường chéo (off-diagonal) vẫn không đổi. Phương pháp này tốn kém về mặt tính toán và kết quả phụ thuộc vào phạm vi các tham số chuẩn hóa doa người dùng cung cấp. Phương pháp thay thế thứ hai của thuật toán chuẩn hóa dựa trên các ước lượng tối ưu của ma trận tương quan [10]. Thuật toán này được gọi là RCCA nhanh (fast RCCA: FRCCA), bởi vì nó tính toán không tốn kém và tương đối nhanh để ước tính kết quả. Trong FRCCA, các hệ số co [10] được ước lượng để nghịch đảo C_{xx}, C_{yy}. Quy trình được sử dụng để thu được ước lượng sai số bình phương tối thiểu của ma trận tương quan có thể được áp dụng để ước tính bất kì ma trận tương quan nào. Phương pháp không giới hạn trong các ma trận tương quan tập nội bộ C_{xx}, C_{yy}; phương pháp này cũng được áp dụng để tìm ước lượng sai số bình phương tối thiểu của C_{xy}. Các hệ số co này làm giảm giá trị của các phần tử ngoài đường chéo của C_{xx}, C_{yy}, trong khi đó các giá trị của các phần tử đường chéo vẫn giữ nguyên. Tuy nhiên, tất cả CCA, RCCA và FRCCA đều mang bản chất không giám sát và không tận dụng được đầy đủ các thông tin về nhãn lớp có sẵn. Để kết hợp thông tin về lớp, một số phiên bản có giám sát của RCCA đã được giới thiệu, được gọi là RCCA có giám sát (supervised RCCA: SRCCA) [10]. Phương pháp này bao gói thông tin nhãn lớp có sẵn để chọn các đặc trưng tương quan tối đa.

Để giải quyết vấn đề kì dị của ma trận hiệp phương sai, RCCA tăng các phần tử đường chéo, trong khi FRCCA làm giảm các phần tử không đường chéo của ma trận hiệp phương sai. Vấn đề này đã được [9] đề xuất một thuật toán trích xuất đặc trưng mới, tích hợp các ưu điểm của cả RCCA và FRCCA để xử lý vấn đề điều kiện không đúng của ma trận hiệp phương sai. Các phần tử đường chéo của ma trận hiệp phương sai được tăng lên bằng cách sử dụng các tham số chuẩn hóa (regularization), trong khi các phần tử ngoài đường chéo bị giảm bằng cách sử dụng các tham số co (shrinkage). Nó cũng tích hợp các giá trị của phương pháp tiếp cận hypercuboid thô để trích xuất các đặc trưng tương quan, liên quan nhất và có ý nghĩa nhất.

a) Khái niệm cơ bản phân tích tương quan chính tắc

CCA thu được hai vector cơ sở định hướng w_x, w_y sao cho hệ số tương quan được tính theo công thức (2.1) lớn nhất, trong đó $\mathbf{C_{xy}} \in \mathbb{R}^{\mathbf{p}\times\mathbf{q}}$ là ma trận hiệp phương sai chéo của X và Y, $\mathbf{C_{xx}} \in \mathbb{R}^{\mathbf{p}\times\mathbf{q}}$ và $\mathbf{C_{yy}} \in \mathbb{R}^{\mathbf{p}\times\mathbf{q}}$ là ma trận hiệp phương sai của X, Y tương ứng. Để tính toán vector cơ sở $\mathbf{w_x}$, $\mathbf{w_y}$, vector riêng của $\sum\sum^{T}$ và $\sum^{T}\sum$ khi ma trận $\sum \in \mathbb{R}^{\mathbf{p}\times\mathbf{q}}$ được định nghĩa theo công thức (2.2):

$$\sum = C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2} \qquad (2.2)$$

Cặp thứ *t* của vector cơ sở được tính theo công thức (2.3):

$$w_{xt} = C_{xx}^{-1/2} \xi_{xt} \text{ và } w_{yt} = C_{yy}^{-1/2} \xi_{yt} \qquad (2.3)$$

Và tập biến chính phương của cặp thứ *t* được tính theo công thức (2.4):

$$u_t = C_{xt}^{T}\text{X và } v_t = C_{yt}^{T}\text{ Y} \qquad (2.4)$$

trong đó ξ_{xt} , ξ_{yt} là giá trị của vector riêng $\sum\sum^{T}$ và $\sum^{T}\sum$ với giá trị riêng ρ_t tương ứng.

b) RCCA với tham số chuẩn hóa và co

Phần này trình bày một thuật toán trích xuất đặc trưng [9], tích hợp một cách khôn ngoan những lợi thế của cả RCCA và FRCCA để xử lý vấn đề kỳ dị của ma trận hiệp phương sai. Phương pháp được đề xuất cũng kết hợp

thông tin tin nhãn lớp có sẵn để làm cho nó có giám sát. Các tham số chuẩn hóa, r_x và r_y biến đổi trong phạm vi $[r_{min}, r_{max}]$, trong đó $r_{min} \leq r_x, r_y \leq r_{max}$. Tập tham số tối ưu r_x và r_y được chọn để cho tương quan Pearson là cực đại, công thức (2.1) được biến đổi thành (2.5) :

$$\mathbf{max}_{\mathbf{r_x r_y}} = \frac{\mathbf{w_x^T C_{xy} w_y}}{\sqrt{\mathbf{w_x^T [C_{xx} + r_x I]\, w_x}}\sqrt{\mathbf{w_y^T [C_{yy} + r_y I] w_y}}} \qquad (2.5)$$

Trong [10], phương pháp FRCCA đã được đề xuất làm cho ma trận hiệp phương sai nghịch đảo được. Ở đây, các tham số co s_x và s_y được sử dụng để xử lý vấn đề kỳ dị của các ma trận hiệp phương sai C_{xx}, C_{yy} tương ứng. Tham số co s_{xy} cũng được sử dụng để tìm ước lượng sai số bình phương trung bình tối thiểu của ma trận hiệp phương sai C_{xy}. Vì thế,

$$[\widetilde{\boldsymbol{C}}_{\boldsymbol{xx}}]_{\boldsymbol{ij}} = (1 - s_x)\,[C_{xx}]_{ij} \text{ và } [\widetilde{\boldsymbol{C}}_{\boldsymbol{yy}}]_{\boldsymbol{ij}} = (1 - s_y)\,[C_{yy}]_{ij}\ , i \neq j$$

$$\text{Và } [\widetilde{\boldsymbol{C}}_{\boldsymbol{xy}}]_{\boldsymbol{ij}} = (1 - s_{xy})\,[C_{xy}]_{ij} \qquad (2.6)$$

Ước tính tốt nhất về các tham số co s_x, s_y và s_{xy} làm cực tiểu hàm nguy cơ của sai số trung bình bình phương, được biểu thị bằng:

$$s_x = \frac{\sum_{\mathbf{i \neq j}} \widehat{\mathbf{V}}\,([\mathbf{Cxx}]\mathbf{ij})}{\sum_{\mathbf{i \neq j}} [\mathbf{C_{xx}^2}]_{\mathbf{ij}}};\ s_y = \frac{\sum_{\mathbf{i \neq j}} \widehat{\mathbf{V}}\,([\mathbf{Cyy}]\mathbf{ij})}{\sum_{\mathbf{i \neq j}} [\mathbf{C_{yy}^2}]_{\mathbf{ij}}};\ s_{xy} = \frac{\sum_{\mathbf{i}}\sum_{\mathbf{i}} \widehat{\mathbf{V}}\,([\mathbf{Cxy}]\mathbf{ij})}{\sum_{\mathbf{i}}\sum_{\mathbf{j}} [\mathbf{C_{xy}^2}]_{\mathbf{ij}}} \qquad (2.7)$$

trong đó, $\widehat{\mathbf{V}}([\mathbf{C_{xx}}]_{\mathbf{ij}})$, $\widehat{\mathbf{V}}([\mathbf{C_{yy}}]_{\mathbf{ij}})$, $\widehat{\mathbf{V}}([\mathbf{C_{xy}}]_{\mathbf{ij}})$ là phương sai thực nghiệm không thiên vị của C_{xx}, C_{yy} và C_{xy} tương ứng. Do đó, để giải quyết vấn đề kì dị này, các ma trận hiệp phương sai và liên hiệp phương sai có thể được xây dựng theo công thức sau:

$$[\tilde{\mathbf{C}}_{\mathbf{xx}}]_{\mathbf{ij}} = \begin{cases} (\mathbf{1} - \mathbf{s_x})[\mathbf{C_{xx}}]_{\mathbf{ij}} & \mathbf{i \neq j} \\ [\mathbf{C_{xx}}]_{\mathbf{ij}} + (\mathbf{r_x} + \mathbf{k d_x}) & \mathbf{i = j} \end{cases} \qquad (2.8)$$

$$[\tilde{\mathbf{C}}_{\mathbf{yy}}]_{\mathbf{ij}} = \begin{cases} (\mathbf{1} - \mathbf{s_y})[\mathbf{C_{yy}}]_{\mathbf{ij}} & \mathbf{i \neq j} \\ [\mathbf{C_{yy}}]_{\mathbf{ij}} + (\mathbf{r_y} + \mathbf{l d_y}) & \mathbf{i = j} \end{cases} \qquad (2.9)$$

$$[\tilde{\mathbf{C}}_{\mathbf{xy}}]_{\mathbf{ij}} = ((\mathbf{1} - \mathbf{s_{xy}})[\mathbf{C_{xy}}]_{\mathbf{ij}}) \quad \text{với mọi i, j} \qquad (2.10)$$

trong đó sự khác biệt phổ biến là d_x, d_y cho r_x và r_y; $k \in \{1,2,\ldots, t_x\}$ và $l \in \{1,2, \ldots, t_y\}$, các tham số t_x, t_y biểu thị số lượng giá trị có thể có của r_x và r_y tương ứng.

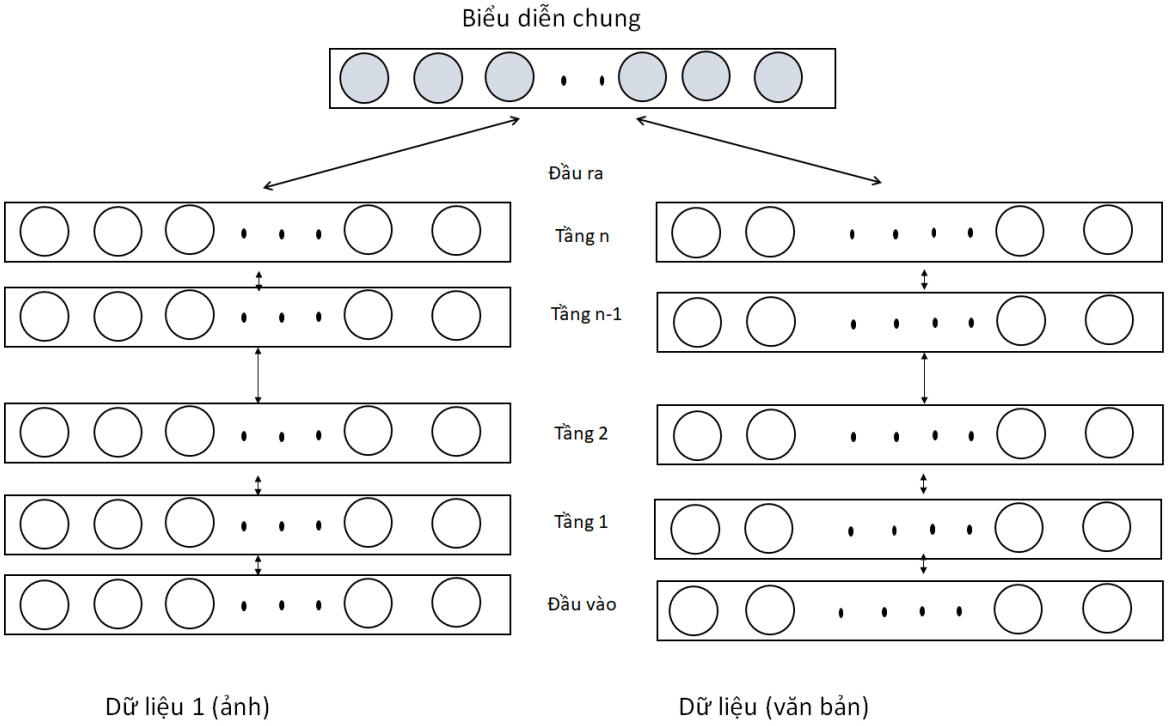
Ngoài CCA, phương pháp bình phương tối thiểu từng phần (Partial Least Squares) PLS cũng được sử dụng cho truy hồi chéo mô hình ảnh, văn bản [16]. Họ sử dụng PLS để chuyển đặc trưng ảnh trong trong không gian văn bản và sau đó học không gian ngữ nghĩa để tính độ tương tự giữa hai mô hình khác nhau theo công thức (2.11):

$$X = T. P^T + E \text{ và } Y = U. Q^T + F \quad (2.11)$$

trong đó, X và Y là ma trận dữ liệu đặc trưng $n \times m$, $n \times p$ tương ứng. T và U là ma trận $n \times l$ chiếu tương ứng trên ma trận nhân tử X, Y. P và Q là ma trận trực giao $m \times l$, $p \times l$ tương ứng. E và F là điều kiện lỗi. Sự phân tách của X và Y được tạo ra để tối đa hóa hiệp phương sai giữa T và U.

2. 2 Phương pháp học sâu

Dữ liệu đa phương thức là các kiểu dữ liệu khác nhau nhưng cùng mô tả cùng sự kiện hoặc chủ đề. Ví dụ, nội dung do người dùng tạo thường chứa nhiều loại dữ liệu khác nhau như ảnh, văn bản và video. Điều này là thách thức lớn với các phương pháp truyền thống là tìm một biểu diễn chung cho nhiều mô hình. Gần đây, sự phát triển học sâu được cộng đồng nghiên cứu được quan tâm và ứng dụng vào giải quyết các bài toán đem lại kết quả hiệu quả hơn so với các phương pháp truyền thống. Học sâu thiết kế nhiều mạng để học các đặc trưng sâu hơn trên các mô hình khác nhau để thu được biểu diễn học hiệu quả [12, 15, 18] đặc biệt cho xử lý ảnh hay truy hồi chéo giữa ảnh và văn bản [6, 14, 21]. Đầu tiên, sử dụng các mô hình mức riêng biệt để học các biểu diễn mức thấp cho mỗi mô hình hay còn gọi là tiền xử lý và trích xuất đặc trưng từ nội dung của dữ liệu đa phương thức, sau đó kết hợp các biểu diễn theo kiến trúc học sâu ở mức độ biểu diễn cao hơn. Trong hình 2.1 minh họa áp dụng học sâu cho tìm biểu diễn chung cho dữ liệu đa phương thức giữa ảnh và văn bản.



Hình 2.1: Minh họa học sâu cho học biểu diễn kết hợp cho ảnh và văn bản

Một nghiên cứu nổi bật học biểu diễn sử dụng học sâu của tác giả Andrew và cộng sự [1], đã đề xuất kỹ thuật phân tích tương quan chính tắc sâu DCCA (Deep Cannonical Correlation Analysis). DCCA học phép chiếu phi tuyến tính (nonlinear) phức tạp cho các phương thức dữ liệu khác nhau sao cho các biểu diễn kết quả là tuyến tính tương quan cao. Nhóm tác giả Goodfellow và cộng sự [6] đề xuất học sâu đối lập và được phát triển cho truy hồi chéo mô hình giữa ảnh và văn bản trong gọi là GAN (Generative Adversarial Nets) [14].

a) Phân tích tương quan chính tắc sâu (DCCA)

DCCA tính toán biểu diễn của dữ liệu đa phương thức (hai khung nhìn tương ứng với 2 kiểu dữ liệu của hai mô hình khác nhau) bằng cách truyền chúng qua nhiều lớp xếp chồng lên nhau của hàm chuyển đổi phi tuyến tính. Đầu vào khung nhìn thứ nhất có c_1 đơn vị (unit) và đầu ra là *o* đơn vị. Kí hiệu $x_1 \in \mathbb{R}^{n_1}$ khung nhìn dữ liệu thứ nhất, đầu ra của tầng thứ nhất cho x_1 là $h_1 = s(W_1^1 x_1 + b_1^1) \in \mathbb{R}^{c_1}$, trong đó $W_1^1 \in \mathbb{R}^{c_1 \times n_1}$ là ma trận trọng số học, $b_1^1 \in \mathbb{R}^{c_1}$ là vector thiên vị (bias) và $s: \mathbb{R} \to \mathbb{R}$ là hàm phi tuyến tính. Đầu ra h_1 sau đó được sử dụng tính toán đầu ra cho tầng tiếp theo như $h_2 = s(W_2^1 h_1 + b_2^1) \in \mathbb{R}^{c_1}$ và thực hiện tới khi biểu diễn cuối cùng $f_1(x_1) = s(W_d^1 h_d + b_d^1)$ được tính toán xong, với *d* là số tầng của mạng. Tương tự tính toán $f_2(x_2)$ với

khung nhìn dữ liệu thứ hai x_2 với bộ tham số W_l^1 và b_l^1 với l là số tầng của mạng. Mục đích là để tham số học kết hợp hai khung nhìn W_l^v và b_l^v để mà độ tương quan corr($f_1(X_1)$, $f_2(X_2)$) là lớn nhất có thể theo công thức (2.12)

$$(\theta_1^*, \theta_2^*) = argmaxcorr_{(\theta_1,\theta_2)} (f_1(X_1; \theta_1), f_2(X_2; \theta_2)) \quad (2.12)$$

Để tìm (θ_1^*, θ_2^*), nhóm tác giả tính toán đạo hàm mục tiêu tương quan được ước tính trên dữ liệu đào tạo. Có $H_1 \in \mathbb{R}^{o\times m}$, $H_2 \in \mathbb{R}^{o\times m}$ là các ma trận cột biểu diễn ở mức trên được tạo ra bởi mô hình học sâu trên hai khung nhìn, với *m* là số lượng dữ liệu mẫu huấn luyện. Có $\bar{H}_1 = H_1 - \frac{1}{m} H_1$ là ma trận dữ liệu trung tâm, tương tự với $\bar{H}_2$ và định nghĩa $\hat{\sum}_{12} = \frac{1}{m-1} \bar{H}_1 \bar{H}_2'$ và $\hat{\sum}_{11} = \frac{1}{m-1} \bar{H}_1 \bar{H}_1' + r_1 I$ với r_1 là hằng số chuẩn, tương tự tính $\hat{\sum}_{22}$. Giả sử rằng, $r_1 > 0$ để $\hat{\sum}_{11}$ không âm. Tổng độ tương quan k thành phần của H_1 và H_2 là tổng của k giá trị riêng của ma trận $T = \hat{\sum}_{11}^{-1/2} \hat{\sum}_{12} \hat{\sum}_{22}^{-1/2}$. Nếu k = o, độ tương quan sẽ được tính theo công thức (2.13):

$$\text{corr}(H_1, H_2) = \|T\| = \text{tr}(T'T)^{1/2} \quad (2.13)$$

Các tham số W_l^v và b_l^v của DCCA được huấn luyện tối ưu bởi sử dụng tối ưu dựa trên đạo hàm. Để tính toán đoạn hàm của độ tương quan *corr(H_1, H_2)* với tất cả các tham số W_l^v và b_l^v có thể đạo hàm với bởi H_1 và H_2 và sau đó dùng lan truyền ngược. Nếu SVD (singular value decomposition) của T định nghĩa là $T = UDV'$, sau đó đạo hàm của *corr(H_1, H_2)* được tính theo công thức (2.14):

$$\frac{\partial\, corr(H1,H2)}{\partial H1} = \frac{1}{m-1}(2. \nabla_{11} \bar{H}_1 + \nabla_{12} \bar{H}_2) \quad (2.14)$$

$$\text{trong đó } \nabla_{12} = \hat{\sum}_{11}^{-1/2} UV' \hat{\sum}_{22}^{-1/2} \quad (2.15)$$

Mỗi một tầng sẽ được tính tổng số lỗi bình phương sao cho là nhỏ nhất cục bộ theo công thức (2.16):

$$l_a (W, b) = \left\|\hat{X} - X\right\|_F^2 + \lambda_a(\|W\|_F^2 + \|b\|_2^2) \quad (2.16)$$

trong đó, $\|.\|_F^2$ là Frobenius norm, λ_a là tham số phạt, $X \in \mathbb{R}^{n\times m}$ là ma trận dữ liệu huấn luyện.

b) GAN

Ý tưởng GAN [14] hoạt động đối lập bởi hai mô hình gọi là mô hình sinh (generative model) và mô hình phân biệt (discriminative model). Trong đó mô hình phân biệt sẽ học để xác định các mẫu là mô hình sinh ra hay là từ phân phối của dữ liệu, trong khi mô hình sinh cố gắng tạo ra các mẫu tương tự dữ liệu mẫu thật. Tưởng tượng rằng mô hình sinh có thể coi là tương tự như một nhóm người làm hàng giả cố gắng tạo ra sản phẩm giả và sử dụng nó mà không bị phát hiện, trong khi mô hình phân biệt tương tự như cảnh sát cố gắng phát hiện được ra hàng giả. Cạnh tranh đối lập buộc cả hai nhóm đều phải cải thiện phương pháp.

Mô hình GAN áp dụng đơn giản nhất khi mô hình là các mạng perceptron nhiều tầng. Để học được phân phối p_g trên dữ liệu *x*, nhóm tác giả định nghĩa trước biến nhiễu đầu vào $p_z(z)$, sau đó biểu diễn ánh xạ sang không gian dữ liệu $G(z; \theta_g)$, ở đây G là hàm có thể phân biệt được biểu diễn bằng perceptron nhiều tầng với các tham số θ_g. Nhóm tác giả định nghĩa $D(x; \theta_d)$ là perceptron nhiều tầng chứa đầu ra, D(x) đại diện cho xác suất rằng x đến từ dữ liệu thực chứ không phải đến từ p_g. GAN huấn luyện D để tối đa hóa xác suất chỉ định nhãn chính xác cho cả mẫu ví dụ huấn luyện và mẫu được sinh ra từ G, đồng thời huấn luyện G để giảm thiểu $\log(1 - D(G(z)))$. Do đó, mô hình D và G được thể hiện cạnh tranh với hàm giá trị V(G, D) theo công thức (2.17):

$$min_G max_D \ V(G, D) = \mathbb{E}_{x \sim p_{data(x)}}[\log D(x) + \mathbb{E}_{z \sim p_{z(z)}}[\log(1 - D(G(z)))]] \quad (2.17)$$

Mô hình sinh G xác định ngầm phân phối xác suất p_g khi phân bố mẫu ví dụ G(z) đạt được khi $z \sim p_z$. Tối ưu toàn cục của $p_g = p_{data}$, nhóm tác giả quan tâm đầu tiên tới tối ưu mô hình phân biệt D cho bất cứ mô hình sinh G nào. Khi G cố định, mô hình phân biệt tối ưu D được tính theo công thức (2.18):

$$D_G^*(x) = \frac{p_{data(x)}}{p_{data(x)} + p_{g(x)}} \quad (2.18)$$

Nếu G và D có đủ khả năng mở rộng, mô hình phân biệt D sẽ cho phép đạt tối ưu khi nhận G và p_g được cập nhật để cải tiến sau đó thì p_g bão hòa tới p_{data}. Công thức (2.6) được viết lại như sau:

$$\mathbb{E}_{x \sim p_{data(x)}}[\log D_G^*(x)] + \mathbb{E}_{x \sim p_g}[\log(1 - D_G^*(x)] \quad (2.19)$$

Ưu điểm của GAN là có lợi thế về mặt thống kế từ mô hình sinh không chỉ được cập nhật trực tiếp với các ví dụ dữ liệu mà còn các đạo hàm qua mô hình phân biệt. Điều này có nghĩa là các thành phần của đầu vào không được sao chép trực tiếp các tham số của mô hình G. Một ưu điểm khác là mạng này có thể biểu diễn được hình thái, góc cạnh của mẫu ảnh đối với dữ liệu ảnh trong khi các phương pháp dựa trên chuỗi Markov thì phân phối không được rõ nét. Nhược điểm chủ yếu của GAN là không có biểu diễn rõ ràng của $p_{g(x)}$ và D phải được đồng bộ tốt với G trong suốt quá trình huấn luyện, cụ thể là G không được huấn luyện quá nhiều mà không cập nhật D.

2. 3 Một số phương pháp khác

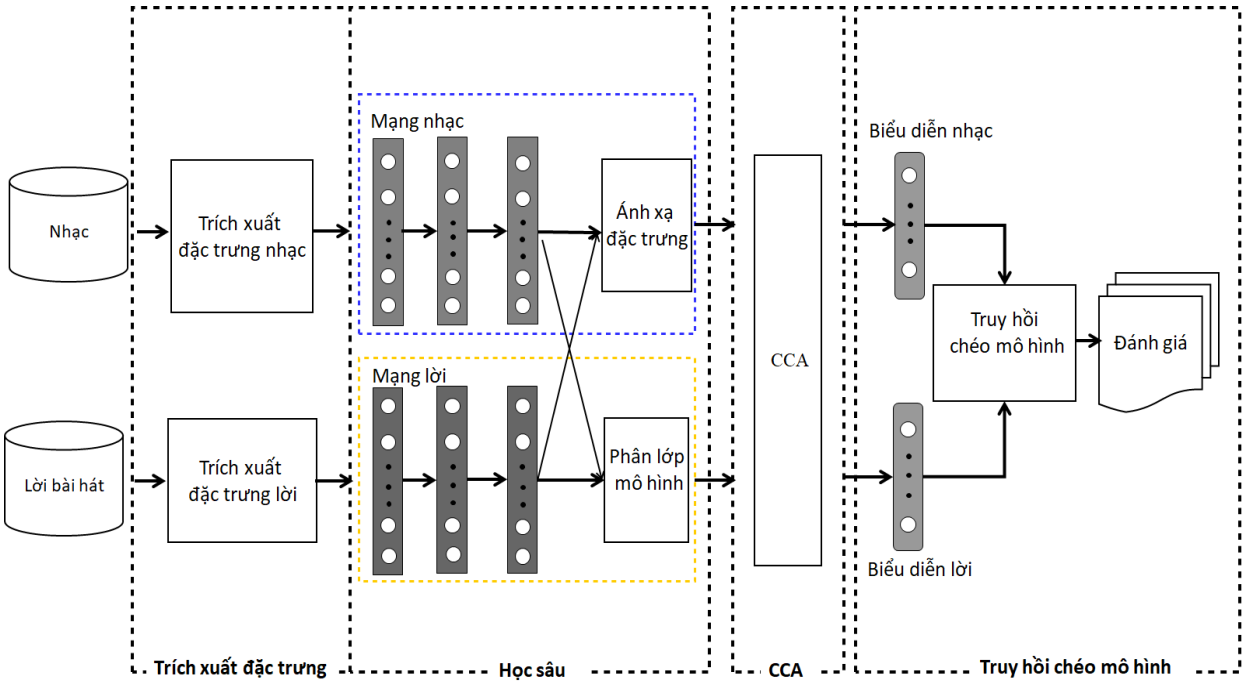
Mô hình chủ đề ẩn được ứng dụng rộng rãi cho bài toán truy hồi chéo mô hình bám giám sát [16]. Để tính toán được sự tương tự giữa ảnh và văn bản mô tả cho ảnh đó, LDA mô hình (latent dirichlet allocation) được mở rộng để học không gian kết nốt chung cho dữ liệu đa phương thức như Corr-LDA (correspondence LDA), tr-mm LDA (topic-regression multi-modal LDA). Corr-LDA sử dụng chủ đề ẩn như các biến ẩn để chia sẻ nơi mà biểu diễn sự tương quan chéo cho dữ liệu đa phương thức. Tr-mm LDA học hai tập riêng biệt của các chủ đề ẩn và mô đun hồi quy nơi mà bắt các hình thức liên kết tổng quát và cho phép một bộ chủ đề được dự đoán tuyến tính từ một chủ đề khác.

Một số kỹ thuật trong phương pháp dựa trên xếp hạng học không gian chung của danh sách các hạng. Yao và cộng sự đề xuất RCCA (ranking canonical correlation analysis) cho truy hồi chéo giữa văn bản và ảnh [16]. RCCA sử dụng điều chỉnh không gian được học bởi CCA để sắp xếp mối quan hệ liên quan giữa các dữ liệu. Trong [16] đề cập nhóm tác giả Lu và cộng sự đề xuất giải thuật xếp hạng chéo mô hình gọi là LSCMR (latent semantic cross-modal ranking). Nhóm tác giả sử dụng SVM để học số liệu sao cho xếp hạng dữ liệu được tạo ra bởi khoảng cách từ một truy vấn có thể được tối ưu hóa so với các độ đo xếp hạng. Tuy nhiên LSCMR không sử dụng hai hướng để xếp hạng, ví dụ như xếp hạng văn bản - truy vấn hình ảnh, xếp hạng hình ảnh – truy vấn văn bản trong huấn luyện.

Chương 3: Mô hình đề xuất

Chương 2 đã trình bày các kiến thức cơ bản về các phương pháp giải quyết bài toán truy hồi chéo mô hình. Để xây dựng được mô hình truy hồi chéo thông tin cho nhạc và lời bài hát, phương pháp tiếp cận trong luận văn xây dựng dựa trên học biểu diễn giá trị thực để tìm ra không gian chung cho phép tính độ tương tự giữa nhạc và lời bài hát. Sử dụng chính nội dung của nhạc và lời bài hát được trích chọn để học biểu diễn cho không gian chung. Sau khi tìm được không gian chung, sử dụng phân tích tương quan chính tắc để chọn ra số lượng thành phần chính tắc phù hợp cho truy hồi chéo mô hình giữa nhạc và lời bài hát.

Truy hồi chéo mô hình cho nhạc và lời bài hát được thực hiện ba pha chính: trích chọn đặc trưng, học biểu diễn chéo mô hình, truy hồi chéo mô hình. Pha thứ nhất trích chọn đặc trưng cho nhạc và lời bài hát cho bước huấn luyện. Pha thứ hai, sử dụng vector đặc trưng qua mạng nơ ron để huấn luyện tìm ra không gian chung cho phép tính toán sự tương tự chéo giữa nhạc và lời bài hát. Áp dụng học sâu cạnh tranh theo [14] để tìm ra không gian biểu diễn chung cho nhạc và lời bài hát. Sau đó sử dụng phân tích tương quan chính tắc để tìm ra số lượng thành phần chính tắc hiệu quả cho việc truy hồi chéo mô hình. Pha thứ ba sử dụng mô hình đề xuất để truy hồi chéo mô hình và đánh giá kết quả của truy hồi chéo mô hình. Luận văn đề xuất mô hình giải quyết bài toán trong hình 3.1.



Hình 3.1: Quy trình truy hồi chéo mô hình cho nhạc và lời bài hát

3.1 Trích chọn đặc trưng

Mỗi bài hát được biểu diễn theo cặp nhạc, lời nhạc và nhãn cảm xúc tương ứng. Mỗi vector đặc trưng âm thanh có 3220 chiều đại diện cho một bản nhạc và mỗi vector đặc trưng lời có 300 chiều đại diện cho lời bài hát.

a) Trích chọn đặc trưng âm thanh

Đối với nhạc, đặc trưng của tín hiệu âm thanh là tham số dùng để phân biệt, nhận dạng các bài hát với nhau. Kích thước toàn bộ tín hiệu âm thanh rất lớn, tín hiệu âm thanh dễ bị biến đổi trong các điều kiện khác nhau nên không thể sử dụng toàn bộ dữ liệu âm thanh của một bài hát làm vector đặc trưng. Do đó, trích chọn đặc trưng tín hiệu âm thanh là vấn đề quan trọng trong các hệ thống xử lý tín hiệu âm thanh nói chung. Cách tiếp cận truyền thống, các vector đặc trưng của tín hiệu âm thanh được xây dựng từ các đặc trưng vật lý của âm thanh như độ to, độ cao, năng lượng, phổ tần số. Trong luận văn trích chọn đặc trưng nhạc, biểu diễn tín hiệu số âm thanh dựa vào tần số Mel – thang đo diễn tả tốt hơn sự nhạy cảm của tai người với âm thanh. Trong nhận dạng tiếng nói, âm thanh nói chung, kỹ thuật trích chọn đặc trưng MFCC (Mel-Frequency Cepstral Coeficients) là phương pháp phổ biến nhất [20]. Kỹ thuật này dựa trên việc thực hiện biến đổi để chuyển dữ liệu âm thành đầu vào đã được biến đổi Fourier cho phổ về thang đo tần số Mel.

Tín hiệu âm thanh được rời rạc hóa bao gồm các mẫu liên tiếp nhau, mỗi mẫu là một giá trị thực, thể hiện giá trị biên độ của âm thanh tại một thời điểm nhất định. Trong luận văn, mỗi bài hát được lấy 30 giây và lấy mẫu với tần số 22050 Hz, mỗi đoạn mẫu với một số lượng nhất định tạo thành một frame. Trích chọn đặc trưng MFCC cho tập đặc trưng mỗi frame. Kết quả là mỗi bài hát sau khi sử dụng kỹ thuật trích chọn đặc trưng MFCC bởi thư viện Librosa [2] sẽ có 646 giá trị đặc trưng cho mỗi một frame và tổng số lượng frame là 20.

b) Trích chọn đặc trưng lời bài hát

Lời bài hát được tiền xử lý tách từ tách câu, loại bỏ nhiễu, lỗi. Các phương pháp trích chọn đặc trưng cho văn bản phổ biến là biểu diễn túi từ (bag of words), túi từ n gram và tính toán mức độ quan trọng của một từ trong tài liệu tf-idf (term frequency – inverse document frequency). Phương pháp túi từ làm mất đi ngữ nghĩa do không quan tâm tới thứ tự của các từ, túi tùi n-gram chỉ xem xét trong ngữ cảnh ngắn và không tốt nếu dữ liệu thưa thớt và số chiều lớn. Phương pháp tf-idf cũng không tốt nếu dữ liệu thưa thớt, khó khăn việc chọn ngưỡng với số chiều nhỏ.

Khắc phục những nhược điểm của các phương pháp trên, Word2vec[3] sử dụng một tập copus qua một mạng nơ ron biểu diễn các từ thành các vector, các vector giữ lại được tính chất ngữ nghĩa. Tức các từ mang ý nghĩa tương tự với nhau thì gần nhau trong không gian vector. Trong xử lý ngôn ngữ tự nhiên, Word2vec là một trong những phương thức của biểu diễn từ (word embedding). Doc2vec [4] không chỉ cho phép biểu diễn từ, câu mà còn cho phép biểu diễn đoạn văn bản. Khi sử dụng Doc2vec mô hình cho phép dễ dàng vector hóa cả một đoạn văn thành một vector có số chiều cố định và nhỏ. Cũng như Word2vec, Doc2vec có hai mô hình là DBOW(Distributed Bag Of Words) và DM (Distributed Memory). Mô hình DBOW không quan tâm thứ tự các từ, huấn luyện nhanh hơn, không sử dụng ngữ cảnh cục bộ. Sau khi huấn luyện xong có các vector biểu diễn của các văn bản. Mô hình DM nối các từ vào tập các từ trong câu. Trong quá trình huấn luyện, vector của từ và đoạn văn đều được cập nhật.

[2] https://librosa.github.io/librosa/
[3] https://radimrehurek.com/gensim/models/word2vec.html
[4] https://radimrehurek.com/gensim/models/doc2vec.html

3.2 Học sâu

Pha thứ hai học sâu áp dụng kiến trúc học sâu dựa trên nghiên cứu của tác giả Wang và cộng sự [14] để tìm ra không gian chung nơi mà các mô hình khác nhau có thể so sánh trực tiếp lẫn nhau dựa trên học đối kháng (adversaial learning). Học đối kháng được thực thi bởi hai quá trình chạy đối lập nhau và cố gắng làm tốt hơn quá trình còn lại. Quá trình thứ nhất ánh xạ đặc trưng (feature projector) coi như pha sinh mẫu (Generative) cố gắng tạo ra một biểu diễn mô hình trong không gian chung và đối kháng lại với pha kia. Quá trình thứ hai phân lớp mô hình (modality classifier) coi như pha phân biệt (Discriminative) cố gắng phân biệt giữa các mô hình khác nhau dựa trên biểu diễn không gian chung. Phương pháp đối kháng học tập đặc trưng nhạc A và lời bài hát T để tìm ra không gian chung $\boldsymbol{S} = \{\boldsymbol{S}_{\mathrm{A}}, \boldsymbol{S}_{\mathrm{T}}\} \in \mathbb{R}^{m \times n}$ cho phép truy hồi chéo mô hình nhạc và lời bài hát. Ở đây hai hàm ánh xạ là $f_{\mathrm{A}}(\mathbf{A}, \theta_A)$, $f_{\mathrm{T}}(\mathbf{T}, \theta_T)$ thực hiện chuyển đổi giá trị đặc trưng của nhạc, lời bài hát tương ứng sang không gian $\boldsymbol{S}$ với cùng số chiều đặc trưng với mạng nơ ron truyền thẳng (feed-forward networks) 3 tầng. Các tầng được kết nối hoàn toàn (fully connected) có các thông số để đảm bảo đủ khả năng biểu diễn giá trị thống kê giữa nhạc và lời bài hát. Sau đó, ánh xạ đặc trưng và phân lớp mô hình được huấn luyện để học đối kháng nhằm mục đích tìm được mô hình phân biệt đặc trưng giữa nhạc và lời dựa trên nhãn.

a) Ánh xạ đặc trưng

Mục tiêu ánh xạ đặc trưng biểu diễn đặc trưng của nhạc và lời nhạc trong không gian biểu diễn mới sao cho nhạc, lời nhạc có thể so sánh trực tiếp về ngữ nghĩa. Ánh xạ đặc trưng gồm hai quá trình: dự đoán nhãn và bảo toàn cấu trúc. Quá trình dự đoán nhãn cho phép chiếu đại diện đặc trưng cho mỗi mô hình trong không gian chung được phân biệt các nhãn ngữ nghĩa. Quá trình bảo toàn cấu trúc đảm bảo rằng các biểu diễn đặc trưng thuộc cùng một nhãn ngữ nghĩa là bất biến trên các mô hình.

Để đảm bảo phân biệt trong mô hình dữ liệu được bảo toàn sau ánh xạ đặc trưng, một phân lớp được thực thi để dự đoán nhãn ngữ nghĩa của các mục được chiếu trong không gian chung. Với mục đích này, mạng truyền thẳng được kích hoạt bởi *softmax* đã được thêm vào đầu mỗi không gian con biểu diễn. Các đặc trưng của mỗi cặp nhạc và lời bài hát đưa vào huấn luyện

bộ phân lớp và đầu ra là phân phối xác suất nhãn ngữ nghĩa mỗi mục. Định nghĩa hàm mất mát phân biệt trong mô hình (intra-modal discrimination loss) kí hiệu $L_{imd}(\theta_{imd})$ như công thức (1), trong đó $\widehat{p_i}$ là xác suất phân phối cho nhạc hoặc lời bài hát, bản chất L_{imd} là hàm loss cross-entropy của phân lớp nhãn trên n cặp nhạc và lời bài hát, θ_{imd} là tham số của bộ phân lớp, y_i là nhãn của mỗi cặp.

$$L_{imd}(\theta_{imd}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i . (log\widehat{p_i}(a_i) + log\widehat{p_i}(t_i))) \quad (1)$$

Quy trình bảo toàn cấu trúc trong mô hình, sử dụng ràng buộc bộ ba để mục tiêu tối thiểu khoảng cách giữa các đại diện của tất cả các mục tương tự ngữ nghĩa từ các mô hình khác, trong khi tối đa khoảng cách giữa các mục khác cùng ngữ nghĩa trong cùng mô hình. Đầu tiên, tất cả các mẫu của các mô hình khác nhưng cùng nhãn được tạo thành các cặp mẫu nhạc và lời bài hát. Nói cách khác, quá trình này xây dựng các cặp có dạng $\{(a_i, t_i^+)\}$ cho cặp có nhãn l_i trong đó lời bài hát với cùng nhãn nhạc được kí hiệu t_i^+ , và dạng $\{(t_i, a_i^+)\}$ cho cặp có nhãn l_i trong đó nhạc với cùng nhãn lời bài hát được kí hiệu là a_i^+. Sau đó, tất cả các khoảng cách giữa các biểu diễn được ánh xạ bởi hai hàm $f_A(\mathbf{A}, \theta_A)$, $f_T(\mathbf{T}, \theta_T)$ trên mỗi cặp mục được tính toán bởi l_2 norm như công thức (3)

$$l_2(a,t) = \| f\mathrm{A}(\mathbf{A}, \theta_A) - f\mathrm{T}(\mathbf{T}, \theta_T) \|_2 \quad (3)$$

Để xây dựng ràng buộc bộ ba, định nghĩa bộ ba với nhãn l_i như sau: $\{(a_i, t_i^+, t_j^-)\}$ và $\{(t_i, a_i^+, a_j^-)\}$. Cuối cùng, tính toán hàm mất mát bất biến trong mô hình kí hiệu là L_{imi} (inter-modal invariance loss) được định nghĩa như trong công thức (4), (5):

$$L_{imi,\ A}(\theta_A) = \sum_{i,j,k}(l_2(a_i, t_j^+) + \lambda . \max(0, \mu - l_2(a_i, t_k^-)) \quad (4)$$

$$L_{imi,\ T}(\theta_T) = \sum_{i,j,k}(l_2(t_i, a_j^+) + \lambda . \max(0, \mu - l_2(t_i, a_k^-)) \quad (5)$$

Kết hợp công thức (4) và (5) được L_{imi} tổng thể cho mô hình nhạc và mô hình lời bài hát $L_{imi,\ A}(\theta_A, \theta_T)$, $L_{imi,\ T}(\theta_A, \theta_T)$ như trong công thức (6):

$$L_{imi}(\theta_A, \theta_T) = L_{imi,\ A}(\theta_A) + L_{imi,\ T}(\theta_T) \quad (6)$$

$$L_{reg} = \sum_{l=1}^{L}(\|W_a^l\|F + \|W_t^l\|F) \quad (7)$$

Trong công thức (7) định nghĩa điều kiện chính quy kí hiệu L_{reg} , F là Frobenius norm – là căn bậc hai của tổng bình phương các phân tử của ma trận và W_a^l , W_t^l đại diện cho các tham số của mạng nơ ron sâu.

Hàm mất mất biểu diễn (loss embedding) cho quy trình ánh xạ đặc trưng kí hiệu là L_{emb} được tính theo công thức (8):

$$L_{emd}(\theta_A, \theta_T, \theta_{imd}) = \alpha . L_{imi} + \beta . L_{imd} + L_{reg} \qquad (8)$$

trong đó hệ số α , β là các tham số điều chỉnh sự đóng góp của L_{imi} và L_{imd} ; còn L_{reg} dùng để ngăn chặn các tham số được học tránh học quá khớp (overfiting learning).

b) Phân lớp mô hình

Phân lớp mô hình định nghĩa một bộ phân lớp *D* với bộ tham số θ_D được coi như hàm phân biệt (Discriminator) trong GAN. Mục tiêu của phân lớp mô hình là phát hiện mô hình nhạc hay lời bài hát khi nhận đầu vào là một vector đặc trưng. Thiết kế mạng học sâu truyền thẳng 3 tầng với bộ tham số θ_D với hàm mất mát đối kháng (adversarial loss) kí hiệu là L_{adv} được định nghĩa trong công thức (9)

$$L_{adv}(\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}(\mathrm{m_i} . (logD(a_i ; \theta_D) + log(1 - D(t_i ; \theta_D))) \qquad (9)$$

trong đó L_{adv} định nghĩa theo hàm mất mát cross-entropy, $\mathrm{m_i}$ là danh sách nhãn của cặp, $D(; \theta_D)$ là xác suất mô hình sinh cho mỗi mục (nhạc hoặc lời bài hát) của mỗi cặp.

c) Tối ưu học đối kháng

Quá trình học biểu diễn đặc trưng tối ưu được thực hiện bằng cách cùng nhau giảm thiểu hàm mất mát L_{emd} công thức (8) và L_{adv} công thức (9). Mục tiêu tối ưu hóa hai quá trình này là đối lập được thể hiện công thức (10), (11):

$$\widehat{\theta_A}, \widehat{\theta_T}, \widehat{\theta_{imd}} = \mathrm{argmin}(L_{emd}(\theta_A, \theta_T, \theta_{imd}) - L_{adv}(\widehat{\theta_D}) \qquad (10)$$

$$\widehat{\theta_D} = \mathrm{argmax}(L_{emd}(\widehat{\theta_A}, \widehat{\theta_T}, \widehat{\theta_{imd}}) - L_{adv}(\theta_D) \qquad (11)$$

Quá trình đối kháng được thực hiện bằng cách sử dụng kỹ thuật tối ưu hóa đạo hàm ngẫu nhiên (stochastic gradient descent optimization algorithm) như kỹ thuật tối ưu hóa đạo hàm ngẫu nhiên Adam. Phương pháp đề xuất chi tiết trong thuật toán 1.

Thuật toán 1: Mã giả cho phương pháp đề xuất

Procedure JointTrain(A, T)
  Trích xuất đặc trưng MFCC cho nhạc, A → F_A
  Trích xuất đặc trưng văn bản cho lời bài hát, T → F_T
  Nhãn cho tập dữ liệu nhạc và lời bài hát, Y = {y_1, y_2, …., y_n}
  for each epoch **do**
    Lấy ngẫu nhiên theo cặp từ F_A , F_T cho batch
    for each batch (w_A, w_T) **do**
      for each pair (a, t) **do**
        Tính toán biểu diễn hàm f_A , f_T
        for k steps **do**
          $\theta_A \leftarrow \theta_A - \mu . \nabla_{\theta_A} (L_{emd} - L_{adv})$ (12)
          $\theta_T \leftarrow \theta_T - \mu . \nabla_{\theta_T} (L_{emd} - L_{adv})$ (13)
          $\theta_{imd} \leftarrow \theta_{imd} - \mu . \nabla_{\theta_{imd}} (L_{emd} - L_{adv})$ (14)
        end for
        $\theta_D \leftarrow \theta_D + \mu . \nabla_{\theta_{imd}} (L_{emd} - L_{adv})$ (15)
        $\boldsymbol{S} = (f_A, f_T)$
        a → **x** by f_A
        t → **y** by f_T
      end for
    end for
  end for
  Chuyển đổi batch (**X**, **Y**)
  Áp dụng CCA cho (**X, Y**) (16)
end Procedure

3.3 Phân tích tương quan chính tắc

Trong thống kê, phân tích tương quan chính tắc (Canonical Correlation Analysis) gọi tắt CCA là một cách suy luận thông tin từ ma trận hiệp phương sai. Nếu có hai vector x và vector y của các biến ngẫu nhiên và có sự tương quan giữa các biến, thì phân tích tương quan chính tắc sẽ tìm được các kết hợp tuyến tính của tập biến x và tập biến y có mối tương quan tối đa với nhau. Phân tích tương quan chính tắc sẽ tạo ra hai biến chính tắc là tổ hợp tuyến tính của các biến trong vector x và vector y. Số lượng biến chính tắc

nhỏ hơn hoặc bằng với số lượng biến trong tập biến nhỏ hơn. Kết quả tương quan chính tắc sẽ cho ta thấy mối quan hệ chặt chẽ hay không chặt chẽ giữa hai vector x và y nhờ vào hệ số tương quan bình phương cho mỗi tập biến.

CCA [4, 9, 10] được dùng để trích xuất đặc trưng ẩn giữa hai tập biến X $\in \mathbb{R}^{p \times n}$ và Y $\epsilon \ \mathbb{R}^{q \times n}$. Ở đây, *n* là số lượng mẫu, p, q là số lượng đặc trưng của X, Y tương ứng. CCA thu được hai vector cơ sở $w_x \in \mathbb{R}^p$ và $w_y \in \mathbb{R}^q$ để tương quan giữa $X^T w_x$ và $Y^T w_y$ là lớn nhất, kí hiệu là ρ, theo công thức (16):

$$\rho = \max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x . w_y^T C_{yy} w_y}} \qquad (16)$$

trong đó $C_{xy} \in \mathbb{R}^{p \times q}$ là ma trận hiệp phương sai chéo của X và Y, $C_{xx} \in \mathbb{R}^{p \times q}$ và $C_{yy} \in \mathbb{R}^{p \times q}$ là ma trận hiệp phương sai của X, Y tương ứng. Để tính toán vector cơ sở w_x, w_y , vector riêng của $\sum \sum^T$ và $\sum^T \sum$ khi ma trận $\sum \in \mathbb{R}^{p \times q}$ được định nghĩa theo công thức (17):

$$\sum = C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2} \qquad (17)$$

Cặp thứ *t* của vector cơ sở được tính theo công thức (18):

$$w_{xt} = C_{xx}^{-1/2} \xi_{xt} \text{ và } w_{yt} = C_{yy}^{-1/2} \xi_{yt} \qquad (18)$$

Và tập biến chính phương của cặp thứ *t* được tính theo công thức (19):

$$u_t = C_{xt}^T \text{X và } v_t = C_{yt}^T \text{Y} \qquad (19)$$

trong đó ξ_{xt} , ξ_{yt} là giá trị của vector riêng $\sum\sum^{\text{T}}$ và $\sum^{\text{T}} \sum$ với giá trị riêng ρ_t tương ứng.

Coi tập biến X là đặc trưng nhạc, tập biến Y là đặc trưng lời đầu vào. Sử dụng phân tích tương quan chính tắc tìm số lượng biến chính tắc cho kết quả tương quan tốt nhất với dữ liệu đa phương thức nhạc và lời bài hát.

3.4 Truy hồi chéo mô hình

Pha truy hồi chéo mô hình sử dụng mô hình học được ở ở pha trước, đầu vào là nhạc hoặc lời bài hát và đầu ra là danh sách các lời bài hát hoặc nhạc liên quan tới truy vấn. Để đánh giá kết quả truy hồi chéo mô hình, luận văn sử dụng độ đo trung bình xếp hạng liên quan MRR (mean reciprocal rank), độ hồi tưởng R (Recall). MRR là một độ đo xem xét vị trí xếp hạng của đối tượng liên quan đầu tiên được trả về được tính theo công thức (20):

$$\text{MRR} = \frac{1}{|\text{N}_\text{q}|} \sum_{i=1}^{\text{N}_\text{q}} \frac{1}{rank_i} \qquad (20)$$

trong đó N_q là tổng số truy vấn được thực hiện, $rank_i$: vị trí xuất hiện đầu tiên của kết quả truy vấn trả về liên quan trong danh sách xếp hạng trả về. Độ hồi tưởng R@k được tính trung bình trên tất cả các truy vấn theo công thức (21):

$$\text{R@k} = \frac{|\text{S}_\text{q} \cap K|}{|\text{S}_\text{q}|} \qquad (21)$$

trong đó S_q là tập các mục (item) liên quan trong cơ sở dữ liệu cho mỗi truy vấn, K là danh sách được xếp hạng của mô hình.

Chương 4: Thực nghiệm và đánh giá

4.1 Dữ liệu và trích xuất đặc trưng

Tập dữ liệu gồm 10.000 cặp nhạc, lời bài hát có 20 nhãn cảm xúc *(giận dữ, hung hăng, trung lập, bình tĩnh, buồn chán, mơ mộng, vui vẻ, lưỡng tính, hạnh phúc, nặng nề, dữ dội, u sầu, vui tươi, yên tĩnh, kỳ quặc, buồn, tình cảm, buồn ngủ, nhẹ nhàng, ngọt ngào*). Mỗi nhãn được thu thập 500 mẫu, mỗi bản nhạc được thu thập trên trang Youtube[5] với các liên kết từ Spotify[6] lấy từ Spotify API, còn lời bài hát của nhạc được thu thập từ trang Musixmatch[7] tương ứng với nhạc. Bảng 4.1 thể hiện chi tiết về dữ liệu và đặc trưng trích xuất.

Bảng 4. 1: Thống kê dữ liệu, đặc trưng và công cụ

Dữ liệu	Số lượng	Đặc trưng	Công cụ
Nhạc	10.000	20 x 161 (MFCCs)	Thư viện: Librosa https://librosa.github.io/librosa/
Lời bài hát	10.000	300 d	Thư viện: Doc2vec https://radimrehurek.com/gensim/models/doc2vec.html

4.2 Môi trường và các công cụ thực nghiệm

Bảng 4.2 chi tiết về môi trường và công cụ thực nghiệm.

Bảng 4. 2: Các công cụ thực nghiệm

STT	Phần mềm	Ý nghĩa	Nguồn
1	Pycharm	Môi trường phát triển	https://www.jetbrains.com/pycharm/
2	Python 2.7	Ngôn ngữ phát triển	https://www.python.org/
3	Tensorflow	Thư viện cho học sâu	https://www.tensorflow.org/
4	Sklearn	Thư viện hỗ trợ các công cụ học máy	http://scikit-learn.org/

[5] https://www.youtube.com/
[6] https://www.spotify.com/
[7] https://www.musixmatch.com/

4.3 Kịch bản thực nghiệm

Luận văn thực hiện 3 kịch bản thực nghiệm: thực nghiệm phương pháp đề xuất, thực nghiệm so sánh với RCCA, thực nghiệm so sánh với các phương pháp khác trong [20] trên cùng một bộ dữ liệu và đánh giá các kết quả thực nghiệm trên các độ đo.

- Thực nghiệm phương pháp đề xuất: thực nghiệm kiểm thử chéo 5 tập (cross-validation) truy hồi chéo mô hình cho nhạc và lời bài hát với các độ đo. Đánh giá kết quả các độ đo trung bình trên 5 tập kiểm tra.
- Thực nghiệm với RCCA: so sánh kết quả thực nghiệm truy hồi chéo mô hình cho nhạc và lời bài hát với CCA. Kiểm thử chéo trên 5 tập và đánh giá kết quả trung bình các độ đo.
- Thực nghiệm so sánh với các phương pháp [20]: PretrainCNN-CCA, DCCA, PretrainCNN-DCCA, JointTrainDCCA cùng bộ dữ liệu để đánh giá. So sánh và đánh giá thực nghiệm với phương pháp đề xuất trong [20] JointTrainDCCA-là phương pháp đề xuất của tác giả Yu và cộng sự [20] đạt kết quả tốt nhất. Mục đích của thực nghiệm so sánh hiệu quả của phương pháp đề xuất với các phương pháp khác.

Thực nghiệm đánh giá thực hiện độ đo MRR trên mức độ thực thể và mức độ nhãn. MRR mức độ thực thể được tính theo công thức (20) dựa trên độ tương tự co-sin mà không quan tâm tới nhãn của nhạc và lời bài hát, kí hiệu là I-MRR-A, I-MRR-L với A, L là sử dụng nhạc, lời bài hát là đầu vào truy vấn tương ứng. MRR mức độ nhãn được tính theo công thức (20) dựa trên nhãn của nhạc và lời bài hát. Thực nghiệm đánh giá với độ đo R@1-A, R@1-L, R@5-A và R@5-L.

4.4 Kết quả thực nghiệm và đánh giá

a) Kết quả thực nghiệm của phương pháp đề xuất

Kết quả thực nghiệm của phương pháp đề xuất khi sử dụng lời nhạc như truy vấn và khi sử dụng nhạc như truy vấn trong Bảng 4. 3. Kết quả các độ đo MRR, độ hồi tưởng khi sử dụng truy vấn là nhạc hay lời bài hát đều cho kết quả xấp xỉ nhau. Điều này chứng tỏ, mô hình đề xuất học ra được không gian chung tốt cho cả nhạc và lời bài hát.

Khi số lượng thành phần chính tắc từ 20 tới 100, kết quả các độ đo tăng từ 20% đến 50%. Điều này chứng tỏ khi không gian chung biểu diễn tốt và

phản ánh đặc trưng chéo mô hình khi tăng số lượng chiều đặc trưng chéo của nhạc hay lời bài hát theo số lượng thành phần chính tắc. Khi thành phần chính tắc là 100 thì kết quả các độ đo đạt từ 40 % đến 50% khi sử dụng truy hồi chéo mô hình cho nhạc hoặc cho lời bài hát.

Bảng 4. 3: Kết quả thực nghiệm của với phương pháp đề xuất

CCA	I-MRR-A	I-MRR-L	C-MRR-A	C-MRR-L	R@1-A	R@1-L	R@5-A	R@5-L
10	0.080	0.081	0.213	0.212	0.045	0.047	0.100	0.099
20	0.200	0.200	0.305	0.305	0.137	0.136	0.251	0.253
30	0.300	0.300	0.387	0.387	0.224	0.224	0.371	0.376
40	0.370	0.366	0.448	0.445	0.288	0.284	0.454	0.447
50	0.415	0.411	0.448	0.484	0.335	0.327	0.498	0.496
60	0.439	0.436	0.506	0.506	0.358	0.354	0.523	0.519
70	0.453	0.449	0.519	0.517	0.371	0.367	0.539	0.535
80	0.456	0.452	0.521	0.519	0.373	0.370	0.540	0.536
90	0.447	0.444	0.515	0.513	0.365	0.362	0.531	0.529
100	0.427	0.425	0.497	0.497	0.349	0.346	0.507	0.505

b) Kết quả thực nghiệm với RCCA

Kết quả thực nghiệm với biến thể RCCA khi sử dụng lời nhạc như truy vấn và khi sử dụng nhạc như truy vấn trong Bảng 4. 4. Tương tự với CCA, RCCA với phương pháp đề xuất truy hồi chéo mô hình hoạt động tốt cho dữ liệu nhạc, lời bài hát với tham số chuẩn hóa *r* được lựa chọn bởi thực nghiệm. Kết quả thực nghiệm RCCA tốt nhất với tham số *r* = 1e-04. Số lượng thành phần chính tắc từ 30 trở đi, kết quả các độ đo tăng từ 20% đến 40%. Khi thành phần chính tắc là 100, các kết quả độ đo khi sử dụng nhạc hoặc lời bài hát truy vấn cũng cho kết quả cao từ 30% đến 40%. Phương pháp đề xuất cho kết quả các độ đo cao hơn so với RCCA từ 5% đến 10% từ 30 thành phần chính tắc trở đi.

Bảng 4. 4: Kết quả thực nghiệm đối với biến thể RCCA

CCA	I-MRR-A	I-MRR-L	C-MRR-A	C-MRR-L	R@1-A	R@1-L	R@5-A	R@5-L
10	0.079	0.084	0.079	0.084	0.052	0.057	0.093	0.099
20	0.163	0.170	0.163	0.170	0.126	0.132	0.190	0.203
30	0.221	0.223	0.221	0.223	0.177	0.179	0.252	0.257
40	0.268	0.263	0.268	0.263	0.221	0.213	0.307	0.308
50	0.295	0.296	0.295	0.296	0.243	0.244	0.343	0.343
60	0.324	0.322	0.324	0.322	0.273	0.265	0.370	0.375
70	0.341	0.343	0.341	0.343	0.288	0.287	0.388	0.394
80	0.357	0.359	0.357	0.359	0.304	0.302	0.409	0.408
90	0.368	0.368	0.368	0.368	0.314	0.310	0.419	0.421
100	0.369	0.371	0.369	0.371	0.317	0.317	0.419	0.417

c) So sánh với các phương pháp khác

Truy hồi chéo mô hình cho nhạc và lời bài hát được nghiên cứu tiên phong bởi tác giả [20] và cộng sự. Luận văn so sánh với phương pháp trong [20]: PretrainCNN-CCA, DCCA, PretrainCNN-DCCA, JointTrainDCCA cùng bộ dữ liệu để đánh giá.

Kịch bản so sánh: thực hiện thực nghiệm so sánh truy hồi chéo mô hình trên các độ đo MRR mức độ thực thể và mức độ nhãn, R@1, R@5 khi sử dụng nhạc hoặc lời truy vấn.

Bảng 4.5 và 4.6 kết quả thực nghiệm so sánh với bốn phương pháp trong [20] trên độ đo MRR mức độ thực thể tương ứng khi sử dụng nhạc, lời bài hát truy vấn. Bảng 4.7 và 4.8 kết quả thực nghiệm so sánh với bốn phương pháp [20] trên độ đo MRR mức độ nhãn tương ứng khi sử dụng nhạc, lời bài hát truy vấn. Bảng 4.9 và 4.10 kết quả thực nghiệm so sánh với JointTrainDCCA phương pháp đạt kết quả cao nhất trong [20] trên độ đo R@1 và R@5 nhãn tương ứng khi sử dụng nhạc, lời bài hát truy vấn.

Bảng 4. 5: Kết quả thực nghiệm so sánh độ đo MRR mức độ thực thể (khi sử dụng nhạc truy vấn)

CCA	PretrainCNN-CCA	DCCA	PretrainCNN-DCCA	JointTrainDCCA	Đề xuất
10	0.022	0.125	0.189	0.247	0.080
20	0.040	0.168	0.225	0.254	0.200
30	0.054	0.183	0.236	0.256	0.300
40	0.069	0.183	0.239	0.256	0.370
50	0.078	0.178	0.237	0.256	0.415
60	0.085	0.177	0.240	0.257	0.439
70	0.090	0.174	0.239	0.256	0.453
80	0.094	0.171	0.237	0.257	0.456
90	0.098	0.164	0.238	0.257	0.447
100	0.099	0.154	0.237	0.257	0.427

Kết quả độ đo MRR mức độ thực thể khi sử dụng nhạc là truy vấn ở Bảng 4.5 của phương pháp đề xuất của luận văn cao hơn so với phương pháp PretrainCNN-CCA, DCCA, PretrainCNN-DCCA, JointTrainDCCA. Kết quả phương pháp đề xuất luận văn với MRR mức độ thực thể từ 40% đến 50% từ thành phần chính tắc 40 trở đi, trong khi PretrainCNN-CCA là 10%, DCCA trung bình là 15%, PretrainCNN-DCCA xấp xỉ 25% và JointTrainDCCA xấp xỉ 25%. So với PretrainCNN-CCA, DCCA, phương phát đề xuất có độ đo MRR cao hơn từ 10% đến 30% từ thành phần chính tắc 30 trở đi. MRR so với PretrainCNN-DCCA, JointTrainDCCA cao hơn từ 5% đến 15% từ thành phần chính tắc 40 trở đi.

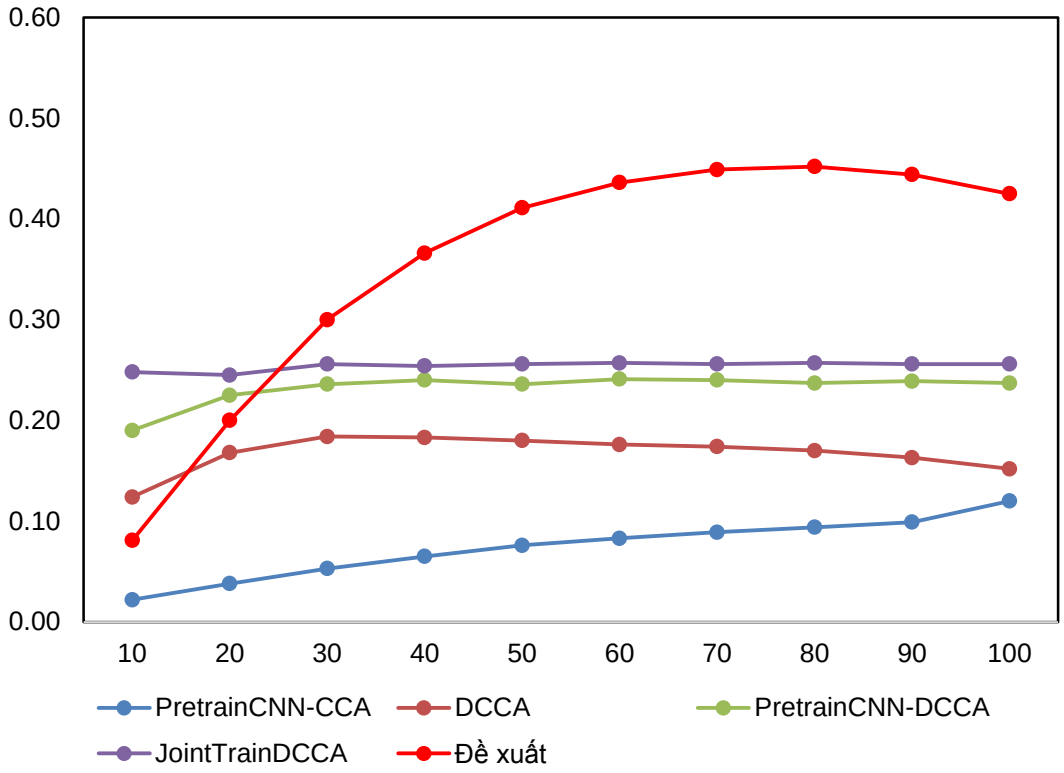
Bảng 4. 6: Kết quả thực nghiệm so sánh độ đo MRR mức độ thực thể (khi sử dụng lời bài hát truy vấn)

CCA	PretrainCNN-CCA	DCCA	PretrainCNN-DCCA	JointTrainDCCA	Đề xuất
10	0.022	0.124	0.190	0.248	0.081
20	0.038	0.168	0.225	0.245	0.200
30	0.053	0.184	0.236	0.256	0.300
40	0.065	0.183	0.240	0.254	0.366
50	0.076	0.180	0.236	0.256	0.411
60	0.083	0.176	0.241	0.257	0.436
70	0.089	0.174	0.240	0.256	0.449
80	0.094	0.170	0.237	0.257	0.452
90	0.099	0.163	0.239	0.256	0.444
100	0.120	0.152	0.237	0.256	0.425

Kết quả độ đo MRR mức độ thực thể khi sử dụng lời bài hát là truy vấn ở Bảng 4.6 của phương pháp đề xuất của luận văn cao hơn so với bốn phương pháp PretrainCNN-CCA, DCCA, PretrainCNN-DCCA, JointTrainDCCA. Kết quả MRR của phương pháp đề xuất luận văn so với bốn phương pháp ở bảng 4.6 khá tương tự với bảng 4.5. So với PretrainCNN-CCA, DCCA, phương phát đề xuất có độ đo MRR cao hơn từ 10% đến 30% từ thành phần chính tắc 30 trở đi. MRR so với PretrainCNN-DCCA, JointTrainDCCA cao hơn từ 5% đến 15% từ thành phần chính tắc 40 trở đi.

Kết quả độ đo MRR mức độ thực thể ở Bảng 4.5 và 4.6 khi sử dụng nhạc hay lời bài hát truy vấn gần như tương tự nhau, chứng tỏ phương pháp đề xuất hoạt động tốt truy hồi chéo mô hình cho nhạc và lời bài hát.

Hình 4.1 So sánh kết quả độ đo MRR mức độ thực thể khi sử dụng nhạc hay lời bài hát truy vấn.



Hình 4. 1: Biểu đồ đường so sánh phương pháp đề xuất với các phương pháp khác trên độ đo MRR mức độ thực thể

Bảng 4. 7: Kết quả thực nghiệm so sánh độ đo MRR mức độ nhãn (khi sử dụng nhạc truy vấn)

CCA	PretrainCNN-CCA	DCCA	PretrainCNN-DCCA	JointTrainDCCA	Đề xuất
10	0.172	0.260	0.313	0.364	0.213
20	0.187	0.296	0.344	0.367	0.305
30	0.199	0.307	0.349	0.368	0.387
40	0.212	0.307	0.356	0.370	0.448
50	0.218	0.304	0.358	0.373	0.448
60	0.225	0.302	0.355	0.370	0.506
70	0.230	0.298	0.358	0.370	0.519
80	0.234	0.294	0.352	0.370	0.521
90	0.235	0.294	0.356	0.370	0.515
100	0.233	0.282	0.354	0.374	0.497

Kết quả độ đo MRR mức độ nhãn khi sử dụng nhạc là truy vấn ở Bảng 4.7 của phương pháp đề xuất của luận văn cao hơn so với phương pháp PretrainCNN-CCA, DCCA, PretrainCNN-DCCA, JointTrainDCCA. Kết quả MRR mức độ nhãn của phương pháp đề xuất luận văn khi sử dụng nhạc là truy vấn từ 38% đến 52% từ thành phần chính tắc 20 trở đi. Từ thành phần chính tắc 10 đến 100, phương pháp đề xuất của luận văn đã cho kết quả MRR cao hơn từ 5% đến 25% đối với PretrainCNN-CCA. Phương pháp đề xuất có MRR cao hơn từ 5% đến 20% đối với DCCA từ thành phần 30 trở đi. So với PretrainCNN-DCCA, JoinTraintDCCA, phương pháp đề xuất cao hơn từ 5% đến 10%.

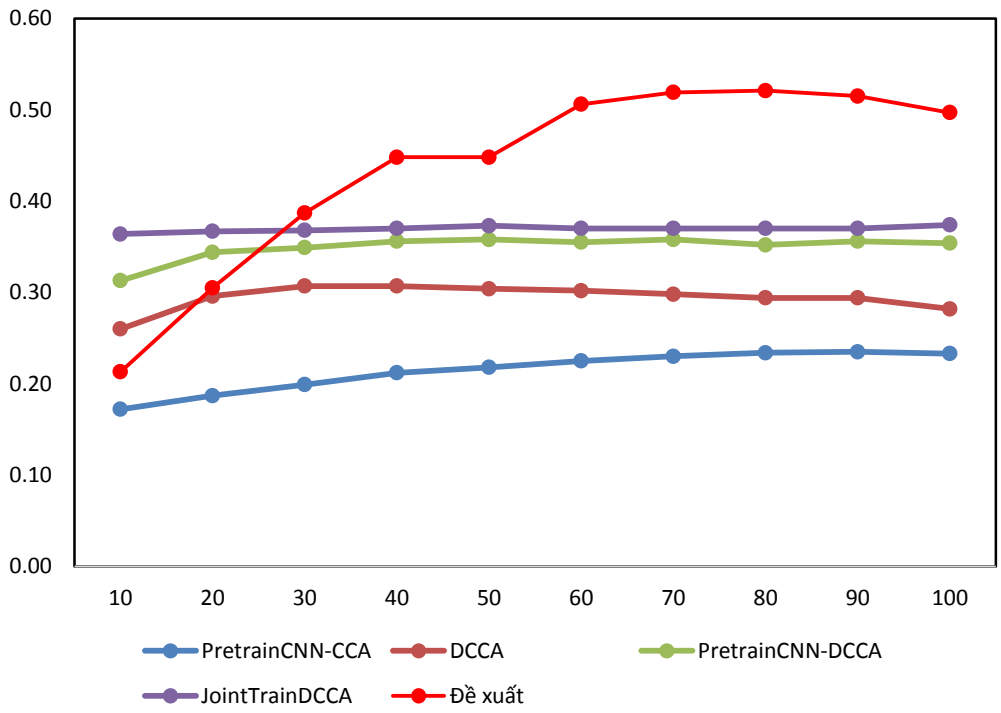
Bảng 4. 8: Kết quả thực nghiệm so sánh độ đo MRR mức độ nhãn (khi sử dụng lời bài hát truy vấn)

CCA	PretrainCNN-CCA	DCCA	PretrainCNN-DCCA	JointTrainDCCA	Đề xuất
10	0.170	0.256	0.314	0.366	0.212
20	0.188	0.294	0.344	0.368	0.305
30	0.198	0.305	0.351	0.372	0.387
40	0.208	0.307	0.358	0.365	0.445
50	0.220	0.306	0.455	0.373	0.484
60	0.223	0.302	0.356	0.374	0.506
70	0.231	0.298	0.360	0.371	0.517
80	0.236	0.290	0.354	0.370	0.519
90	0.237	0.288	0.356	0.369	0.513
100	0.238	0.280	0.355	0.375	0.497

Kết quả độ đo MRR mức độ nhãn khi sử dụng lời bài hát là truy vấn ở Bảng 4.8 của phương pháp đề xuất của luận văn cao hơn so với phương pháp PretrainCNN-CCA, DCCA, PretrainCNN-DCCA, JointTrainDCCA. Kết quả MRR mức độ nhãn của phương pháp đề xuất khi sử dụng lời bài hát là truy vấn từ 38% đến 52% từ thành phần chính tắc 20 trở đi. Từ thành phần chính tắc 10 đến 100, phương pháp đề xuất của luận văn đã cho kết quả MRR cao hơn từ 5% đến 25% đối với PretrainCNN-CCA. Phương pháp đề xuất có MRR cao hơn từ 5% đến 20% đối với DCCA từ thành phần 30 trở đi. So với PretrainCNN-DCCA, JoinTraintDCCA, phương pháp đề xuất cao hơn từ 5% đến 10%.

Kết quả MRR mức độ nhãn ở bảng 4.7 và 4.8 khá tương tự nhau, chứng tỏ mô hình đề xuất hoạt động hiệu quả cho cả nhạc lẫn lời bài hát khi truy vấn.

Hình 4.2 So sánh kết quả độ đo MRR mức độ nhãn khi sử dụng nhạc hay lời bài hát truy vấn



Hình 4. 2: Biểu đồ đường so sánh phương pháp đề xuất với các phương pháp khác trên độ đo MRR mức độ nhãn

Bảng 4. 9: Kết quả độ đo hồi tưởng khi so sánh với JointTrainDCCA (khi sử dụng nhạc truy vấn)

CCA	**R@1** **JointTrain DCCA**	**R@1** **Đề xuất**	**R@5** **JointTrain DCCA**	**R@5** **Đề xuất**
10	0.233	0.045	0.257	0.100
20	0.243	0.137	0.262	0.251
30	0.245	0.224	0.263	0.371
40	0.245	0.288	0.262	0.454
50	0.246	0.335	0.262	0.498
60	0.246	0.358	0.263	0.523
70	0.246	0.371	0.263	0.539
80	0.246	0.373	0.264	0.540
90	0.247	0.365	0.263	0.531
100	0.246	0.349	0.263	0.507

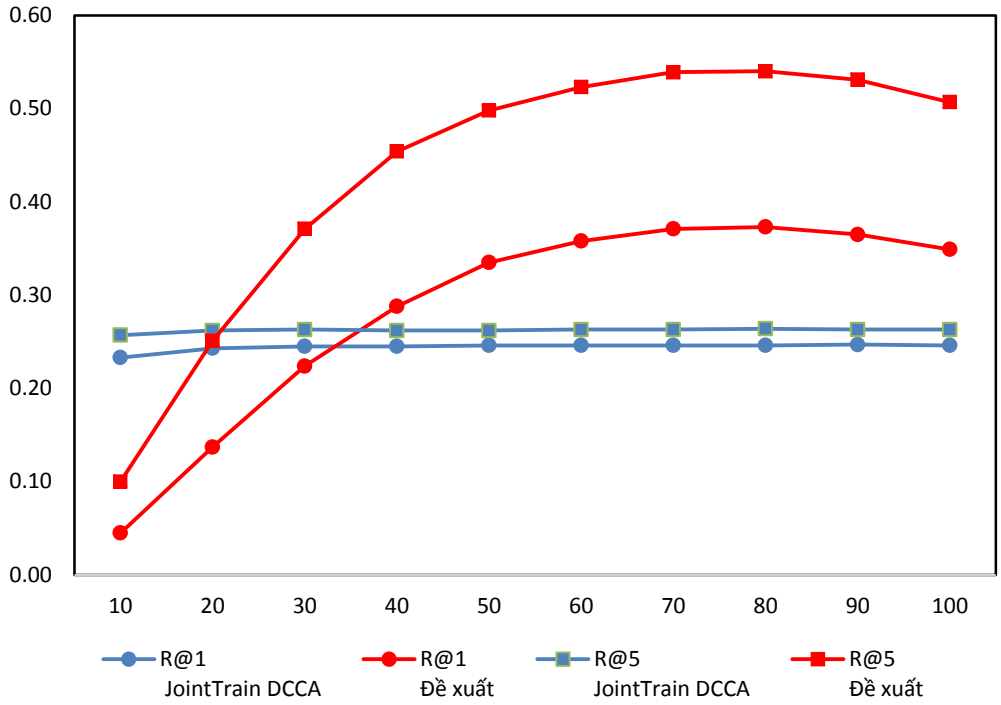
Kết quả độ đo R@ 1 và R@5 trên khi sử dụng nhạc là truy vấn ở Bảng 4.9 của phương pháp đề xuất luận văn cao hơn so với phương pháp JointTrainDCCA. Kết quả R@ 1 của phương pháp đề xuất luận văn khi sử dụng nhạc là truy vấn từ 25% đến 35% từ thành phần chính tắc 40 trở đi cao hơn từ 5% đến 10% so với phương pháp JointTrainDCCA. . Kết quả R@ 5 của phương pháp đề xuất luận văn khi sử dụng nhạc là truy vấn từ 25% đến 54% từ thành phần chính tắc 30 trở đi cao hơn từ 10% đến 25% so với phương pháp JointTrainDCCA

Bảng 4. 10: Kết quả độ đo hồi tưởng khi so sánh với JointTrainDCCA (khi sử dụng lời bài hát truy vấn)

CCA	R1 JointTrain DCCA	R1 Đề xuất	R5 JointTrain DCCA	R5 Đề xuất
10	0.235	0.047	0.257	0.099
20	0.242	0.136	0.261	0.253
30	0.245	0.224	0.263	0.376
40	0.244	0.284	0.261	0.447
50	0.246	0.327	0.262	0.496
60	0.247	0.354	0.263	0.519
70	0.245	0.367	0.263	0.535
80	0.247	0.370	0.264	0.536
90	0.246	0.362	0.263	0.529
100	0.247	0.346	0.262	0.505

Kết quả trên độ đo R@ 1 và R@5 khi sử dụng lời bài hát là truy vấn Bảng 4.10 chỉ ra rằng phương pháp đề xuất của luận văn hoạt động tốt so với phương pháp JointTrainDCCA. Kết quả R@ 1 của phương pháp đề xuất luận văn khi sử dụng lời bài hát là truy vấn từ 25% đến 35% từ thành phần chính tắc 40 trở đi cao hơn từ 5% đến 10% so với phương pháp JointTrainDCCA. Kết quả R@ 5 của phương pháp đề xuất luận văn khi sử dụng lời bài hát là truy vấn từ 25% đến 50% từ thành phần chính tắc 30 trở đi cao hơn từ 10% đến 25% so với phương pháp JointTrainDCCA.

Hình 4.3 so sánh kết quả độ đo R@1 và R@5 của phương pháp đề xuất với JointTrainDCCA [20].



Hình 4. 3 : Biểu đồ đường so sánh phương pháp đề xuất với các phương pháp khác trên độ đo R@1 và R@5

KẾT LUẬN

Truy hồi chéo mô hình không chỉ là chủ đề quan tâm của cộng đồng nghiên cứu thế giới mà còn nhận sự quan tâm của công nghiệp. Các nghiên cứu và ứng dụng nhằm cải tiến và đáp ứng được nhu cầu truy vấn chéo thông tin giữa các dữ liệu đa phương thức của người dùng. Cùng góp phần vào trào lưu nghiên cứu thế giới, luận văn có tên đề tài truy hồi chéo mô hình cho nhạc và lời bài hát thực hiện để xây dựng mô hình cho phép truy hồi chéo khi sử dụng nhạc là truy vấn hoặc khi sử dụng lời bài hát là truy vấn. Luận văn đề xuất ra phương pháp mới kết hợp bởi học sâu và phân tích tương quan chính tắc và sử dụng mô hình đề xuất để truy hồi chéo cho nhạc và lời bài hát. Đồng thời luận văn cũng đánh giá và so sánh hiệu quả của phương pháp đề xuất với các phương pháp điển hình khác để chứng minh phương pháp đề xuất khả quan để ứng dụng vào thực tiễn. Kết quả phương pháp đề xuất cao hơn so với các phương pháp so sánh trên cùng một tập dữ liệu. Kết quả độ đo MRR, R@1, R@5 của phương pháp đề xuất trong luận văn khi sử dụng nhạc hay sử dụng lời bài hát truy vấn từ 30% đến 50% trên tập dữ liệu âm nhạc. Phương pháp đề xuất trong luận văn có thể được ứng dụng cho các hệ thống tìm kiếm chéo trên các trang âm nhạc nhằm đáp ứng nhu cầu truy vấn của người dùng.

TÀI LIỆU THAM KHẢO

1. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: *Deep canonical correlation analysis*. In International Conference on Machine Learning. pp. 1247-1255 (2013)
2. Boutell, M., Luo, J.: *Photo classification by integrating image content and camera metadata*. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. vol. 4, pp. 901-904. IEEE (2004)
3. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: *Multi-view clustering via canonical correlation analysis*. In Proceedings of the 26th annual international conference on machine learning. pp. 129-136. ACM (2009)
4. De Bie, T., De Moor, B.: *On the regularization of canonical correlation analysis*. Int. Sympos. ICA and BSS pp. 785-790 (2003)
5. Feng, F., Li, R., Wang, X.: *Deep correspondence restricted boltzmann machine for cross-modal retrieval*. Neurocomputing **154**, 50-60 (2015)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: *Generative adversarial nets*. In: Advances in neural information processing systems. pp. 2672-2680 (2014)
7. Hu, X., Downie, J.S., Ehmann, A.F.: *Lyric text mining in music mood classification*. American music **183**(5,049), 2-209 (2009).
8. Le, Q., Mikolov, T.: *Distributed representations of sentences and documents*. In International Conference on Machine Learning. pp. 1188-1196 (2014)
9. Mandal, A., Maji, P.: *Regularization and shrinkage in rough set based canonical correlation analysis*. In International Joint Conference on Rough Sets. pp. 432-446. Springer (2017)
10. Mandal, A., Maji, P.: *Faroc: fast and robust supervised canonical correlation analysis for multimodal omics data*. IEEE transactions on cybernetics 48(4), 1229-1241 (2018)
11. McAuley, J., Leskovec, J.: *Image labeling on a network: using social-network metadata for image classification*. In European conference on computer vision. pp. 828-841. Springer (2012)
12. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: *Multimodal deep learning*. In Proceedings of the 28th international conference on machine learning (ICML-11). pp. 689-696 (2011)

13. Peng, Y., Huang, X., Qi, J.: *Cross-media shared representation by hierarchical learning with multiple deep networks*. In IJCAI. pp. 3846-3853 (2016)
14. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: *Adversarial cross-modal retrieval*. In Proceedings of the 2017 ACM on Multimedia Conference. pp. 154-162. ACM (2017)
15. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: *Learning coupled feature spaces for cross-modal matching*. In Proceedings of the IEEE International Conference on Computer Vision. pp. 2088-2095 (2013)
16. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: *A comprehensive survey on cross-modal retrieval*. arXiv preprint arXiv:1607.06215 (2016)
17. Xia, R., Pan, Y., Lai, H., Liu, C., Yan, S.: *Supervised hashing for image retrieval via image representation learning*. In AAAI. vol. 1, p. 2 (2014)
18. Yan, F., Mikolajczyk, K.: *Deep correlation for matching images and text*. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3441-3450 (2015)
19. Yao, T., Mei, T., Ngo, C.W.: *Learning query and image similarities with ranking canonical correlation analysis*. In Proceedings of the IEEE International Conference on Computer Vision. pp. 28-36 (2015)
20. Yu, Y., Tang, S., Raposo, F., Chen, L.: *Deep cross-modal correlation learning for audio and lyrics in music retrieval*. arXiv preprint arXiv:1711.08976 (2017)
21. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: *Text to photo-realistic image synthesis with stacked generative adversarial networks*. arXiv preprint (2017)
22. Zhang, J., Peng, Y., Yuan, M.: *Unsupervised generative adversarial cross-modal hashing*. arXiv preprint arXiv:1712.00358 (2017)