# ĐẠI HỌC BÁCH KHOA HÀ NỘI

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

A red and white rectangular sign with a yellow and red design

Description automatically generated

**BÁO CÁO**

## HỌC PHẦN PROJECT III

*Đề tài: Xây dựng hệ gợi ý tin tức( News Recommendation System)*

**Giảng viên hướng dẫn:** TS. Nguyễn Thị Thu Hương

**Khoa:** Khoa học máy tính

**Trường:** Công nghệ Thông tin và Truyền thông

**Sinh viên:** Trần Trọng Khang 20204660

*­*

*Hà Nội, Tháng 01/2024*

**LỜI CẢM ƠN**

Em xin được gửi lời cảm ơn đến TS. Nguyễn Thị Thu Hương đã hỗ trợ em trong quá trình thực hiện bài tập lớn học phần Project III, giúp em hoàn thành nhiệm vụ của học phần này. Với kiến thức có hạn và thời gian hạn chế, báo cáo cùng với bài tập lớn của em không thể tránh khỏi sai sót. Em mong cô quan tâm và góp ý để bài tập lớn của em được hoàn thiện hơn.

**MỤC LỤC**

[ĐẠI HỌC BÁCH KHOA HÀ NỘI 1](#_Toc157692974)

[HỌC PHẦN PROJECT III 1](#_Toc157692975)

[**I.** **Tổng quan đề tài** 4](#_Toc157692976)

[**1.1.** **Mô tả bài toán** 4](#_Toc157692977)

[**1.2.** **Sơ lược ý tưởng giải quyết bài toán** 4](#_Toc157692978)

[**II.** **Nền tảng lý thuyết** 5](#_Toc157692979)

[**2.1.** **Neural Recommendation with Attentive Multi-View Learning (NAML)** 5](#_Toc157692980)

[**2.2.** **Tổng quan mô hình NRMS (Neural News Recommendation with Multi-Head Self-Attention)** 6](#_Toc157692981)

[**III.** **THỰC NGHIỆM PHƯƠNG PHÁP** 9](#_Toc157692982)

[**3.1.** **Dữ liệu** 9](#_Toc157692983)

[**IV.** **CÁC PHƯƠNG PHÁP ĐÁNH GIÁ** 20](#_Toc157692984)

[**4.1.** **Các tham số đánh giá** 20](#_Toc157692985)

[**4.2.** **Môi trường thực nghiệm** 22](#_Toc157692986)

[**4.3.** **Kết quả huấn luyện** 22](#_Toc157692987)

[**V.** **KẾT LUẬN** 23](#_Toc157692988)

1. **Tổng quan đề tài**
   1. **Mô tả bài toán**

Bài toán gợi ý tin tức từ tập dữ liệu MIND (MS News).

Chi tiết về cuộc thi MIND:

<https://competitions.codalab.org/competitions/24122?fbclid=IwAR2SdbmaGQbVt>

* 1. **Sơ lược ý tưởng giải quyết bài toán**

Em hướng đến xây dựng mô hình gợi ý tin tức với các mục tiêu như sau:

1. Mô hình có thể gợi ý được một danh sách các bài báo cho người đọc với độ đo đánh giá ở mức tốt nhất. Cụ thể về các độ đo đánh giá sẽ được trình bày ở chương 4.

2. Tốc độ huấn luyện và xử lý của mô hình phải đủ nhanh để đáp ứng được giới hạn về phần cứng.

3. Mô hình được huấn luyện bởi một tập dữ liệu đa dạng, đủ lớn.

Để có thể đáp ứng được mục tiêu và giải quyết được các vấn đề nêu trên, trước hết em đề xuất sử dụng bộ dữ liệu tin tức MIND-small do Microsoft cung cấp. Sau đó, em thực hiện huấn luyện các mô hình NAML, NRMS trên MIND-large và sử dụng các độ đo đánh giá như AUC, MRR, nDCG@5, nDCG@10 để đánh giá hiệu năng mô hình.

1. **Nền tảng lý thuyết**
   1. **Neural Recommendation with Attentive Multi-View Learning (NAML)**

NAML là một mô hình gợi ý tin tức sử dụng cơ chế học tập chú ý đa góc nhìn (AĴentive Multi-View Learning), được giới thiệu bởi Chuhan Wu và các cộng sự. Mô hình học biểu diễn tin tức và người dùng bằng cách tận dụng các loại thông tin khác nhau. Mô hình gồm hai thành phần cốt lõi, được mô tả sau đây.

**Bộ mã hóa tin tức (News Encoder)**

Bộ mã hóa tin tức được sử dụng để học biểu diễn của tin tức từ các loại thông tin như tiêu đề (title), nội dung (bodies) và danh mục chủ đề (topic categories) của tin tức. Vì các tin tức khác nhau có đặc trưng về các thông tin tin tức là khác nhau, nên thay vì gộp chúng thành một văn bản dài để biểu diễn tin tức, nhóm tác giả đề xuất một khung học tập chú ý đa góc nhìn để học biểu diễn tin tức một cách thống nhất bằng cách xem xét mỗi loại thông tin như một góc nhìn cụ thể của tin tức.

A diagram of a computer

Description automatically generated

Trong đó, bốn thành phần chính trong bộ mã hóa tin tức là:

1. Bộ mã hóa tiêu đề (Title Encoder)

2. Bộ mã hóa danh mục/thể loại tin tức (Category Encoder)

3. Bộ mã hóa nội dung tin tức (Body Encoder)

4. Lớp gộp-chú ý mức góc nhìn (attention pooling)

Trong bộ mã hóa tiêu đề và nội dung tin tức có hai thành phần quan trọng, là lớp tích chập và lớp gộp chú ý mức từ. Cụ thể, lớp tích chập cho phép học biểu diễn từ vựng ngữ cảnh bằng cách nắm bắt ngữ cảnh cục bộ của chúng, vì ngữ cảnh cục bộ của một số từ là quan trọng để nắm bắt được nội dung. Lớp gộp chú ý mức từ được sử dụng sau lớp tích chập đánh trọng số chú ý cho mỗi từ trong nội dung (biểu thị mức độ quan trọng của từ đó), cho phép chọn lọc các từ quan trọng trong ngữ cảnh của tiêu đề hay nội dung tin tức. Việc nhận biết các từ quan trọng mang lại tiềm năng học biểu diễn tin tức hữu ích hơn.

Ngoài ra, các loại thông tin tin tức khác nhau có thể cho thông tin khác nhau khi học biểu diễn của các tin tức. Một số tiêu đề chứa đựng rất nhiều thông tin của tin tức, tuy nhiên có một số tiêu đề lại ngắn, mơ hồ, không chứa nhiều thông tin. Như vậy, những tiêu đề mang lại nhiều thông tin nên được gắn trọng số cao, còn những tiêu đề mơ hồ nên được gắn trọng số thấp so với nội dung và danh mục chủ đề. Do đó, lớp gộp chú ý mức góc nhìn được sử dụng ở trên cùng bộ mã hóa để mô hình hóa sự hữu ích của các loại thông tin tin tức khác nhau trong việc học biểu diễn tin tức.

**Bộ mã hóa người dùng (User Encoder)**

Bộ mã hóa người dùng dùng để học biểu diễn của người dùng bằng biểu diễn của những tin tức mà họ đã xem. Các tin tức khác nhau được xem bởi cùng một người dùng mang lại thông tin khác nhau để biểu diễn người dùng này. Một số tin tức như “Kết quả trận đấu bóng đá của Manchester United đêm qua” hay “Chương trình The New Mentor đã tìm ra được quán quân” là rất hữu ích để mô hình sở thích của người dùng vì thường xuyên được xem bởi những người quan tâm đến thể thao, giải trí, nhưng tin tức “Dự báo thời tiết tuần tới” ít hữu ích hơn vì nó được nhấp bởi nhiều người dùng. Do đó, trong bộ mã hóa người dùng, nhóm tác giả áp dụng một mạng chú ý tin tức (news aĴention) để học biểu diễn người dùng hữu ích hơn bằng cách chọn lọc tin tức quan trọng.

**Bộ dự đoán khả năng nhấp vào tin tức (Click Predictor)**

Bộ dự đoán khả năng nhấp vào tin tức được sử dụng để dự đoán xác suất người dùng xem một tin tức ứng viên dựa trên biểu diễn của chúng, được tính bằng tích vô hướng của các vec-tơ biểu diễn của người dùng và vec-tơ biểu diễn tin tức ứng cử viên. Xác suất này là kết quả đầu ra cho mô hình.

* 1. **Tổng quan mô hình NRMS (Neural News Recommendation with Multi-Head Self-Attention)**

NRMS là một mô hình gợi ý tin tức sử dụng cơ chế tự chú ý đa đầu (Multi-Head Self-AĴention), được giới thiệu bởi Chuhan Wu và các cộng sự. Ý tưởng của bài báo xuất phát từ: i) tương tác giữa các từ trong tiêu đề tin tức là quan trọng để hiểu rõ tin tức, ii) những tin tức khác nhau được duyệt bởi cùng một người dùng có thể có mối quan hệ với nhau, iii) các từ khác nhau có thể có tầm quan trọng khác nhau trong việc đại diện cho tin tức và iv) những tin tức khác nhau mà cùng một người dùng đọc cũng có thể có độ quan trọng khác nhau trong việc biểu diễn người dùng. Hình 2.6 kiến trúc tổng quan của mô hình. Mô hình bao gồm các thành phần: Bộ mã hóa người dùng (User Encoder), Bộ mã hóa tin tức (News Encoder) và Bộ dự đoán khả năng nhấp vào tin tức (Click Predictor).

A diagram of a block diagram

Description automatically generated

**Bộ mã hóa tin tức (News Encoder)**

Phần đầu tiên là lớp nhúng từ (Word Embedding Layer), nhiệm vụ của nó là chuyển đổi các từ trong một tiêu đề tin tức, giả sử chiều dài là 𝑀, thành các vec-tơ nhúng từ (Word Embedding). Phương pháp được sử dụng trong mô hình NRMS mà các tác giả giới thiệu là embedding GloVe. Về mặt ký hiệu, lớp này chuyển đổi tiêu đề tin tức thành vec-tơ nhúng từ .

Phần thứ hai là mạng tự chú ý đa đầu ở mức từ (Word-Level Multi-Head Self-AĴention) để xem xét giả định mà nhóm tác giả đưa ra - tương tác giữa các từ trong tiêu đề tin tức quan trọng đối với biểu diễn vec-tơ tin tức. Mô hình sử dụng cơ chế tự chú ý đa đầu để có được các vec-tơ nhúng có ngữ cảnh.

Gọi là điểm (score) chú ý tương tác giữa từ thứ 𝑖 và từ thứ 𝑗 trên đầu (head) và là các tham số chiếu (projection parameters) của đầu thứ 𝑘. Biểu diễn của từ thứ 𝑖, ký hiệu, được học trên đầu chú ý thứ 𝑘 được tính như sau:

 



Biểu diễn đa đầu của từ thứ 𝑖 là nối các biểu diễn từ đó trên ℎ đầu:



Phần thứ ba là mạng chú ý cộng mức từ (Additive Word Attention), tương ứng với giả định các từ khác nhau trong tiêu đề tin tức có sự quan trọng khác nhau. Cơ chế chú ý ở phần này cho phép lựa chọn những từ quan trọng trong tiêu đề tin tức để học các biểu diễn tin tức. Trọng số chú ý của từ thứ 𝑖 trong tiêu đề tin tức được tính bởi công thức:





trong đó, và  là các tham số chiếu,  là vec-tơ query(. Biểu diễn cuối cùng của một tin tức là tổng trọng số của biểu diễn các từ, hay:



**Bộ mã hóa người dùng (User Encoder)**

Bộ mã hóa người dùng được sử dụng để học các biểu diễn của người dùng từ các tin tức họ đã xem, gồm hai thành phần.

Phần đầu tiên là một mạng tự chú ý đa đầu mức tin tức. Thông thường, các bài báo tin tức mà cùng một người dùng đã xem có thể có một mức độ liên quan. Ngoài ra, một bài báo tin tức có thể tương tác với nhiều bài báo tin tức khác mà cùng một người dùng đã xem. Do đó, nhóm tác giả đề xuất áp dụng cơ chế tự chú ý đa đầu tiên để cải thiện các biểu diễn của tin tức bằng cách bắt kịp các tương tác (interaction) của chúng. Giả sử ta lấy lịch sử với 𝑁 tin tức, được biểu diễn bởi. Biểu diễn của tin tức thứ 𝑖 được học trên đầu tự chú ý thứ 𝑘 được tính như sau:





trong đó, đại diện cho sự quan trọng tương đối của tương tác giữa tin thứ 𝑖 và tin thứ 𝑗 trên đầu 𝑘, và là các tham số (projection parameters) của đầu tự chú ý tin tức thứ 𝑘. Cuối cùng, nối các biểu diễn đầu ra trên các đầu tự chú ý để tạo ra biểu diễn cho tin tức thứ 𝑖:



Phần thứ hai là mạng chú ý cộng mức tin tức (Additive News Attention), tương ứng với giả định các tin tức khác nhau có thể mang lại các thông tin khác nhau khi đại diện cho người dùng. Theo đó, trọng số chú ý  của tin tức thứ 𝑖 được tính bởi công thức:





trong đó, và  là các tham số trong mạng chú ý,  là vec-tơ truy vấn (query vec-tơ). Biểu diễn cuối cùng của các tin tức được xem bởi người dùng là tổng trọng số của biểu diễn các tin tức, hay:



**Bộ dự đoán khả năng nhấp vào tin tức (Click Predictor)**

Ký hiệu cho biểu diễn của một tin tức ứng cử viên  là  . Tương tự so với mô hình NAML, xác suất nhấp vào tin tức của người dùng là kết quả đầu ra của mô hình, tính bởi công thức:



So với mô hình NAML, mô hình NRMS tập trung tăng cường việc mô hình hóa sở thích người dùng.

1. **THỰC NGHIỆM PHƯƠNG PHÁP**
   1. **Dữ liệu**

Bộ dữ liệu MIND (Microsoft News Dataset) cho gợi ý tin tức được thu thập từ nhật ký hành vi ẩn danh của trang web Microsoft News. Microsoft đã lấy mẫu ngẫu nhiên một triệu người dùng có ít nhất năm lần nhấp vào tin tức trong sáu tuần, từ ngày 12 tháng 10 đến ngày 22 tháng 11 năm 2019.

A diagram of a test

Description automatically generated with medium confidence

Bộ dữ liệu được sử dụng trong dự án là MIND-large. MIND-large gồm ba thư mục: MINDlarge-train chứa dữ liệu training, MINDlarge-dev chứa dữ liệu validation và MINDlarge-test chứa dữ liệu tesing. Cấu trúc của từng thư mục như sau:

1. Tệp behaviors.tsv chứa nhật ký người dùng

2. Tệp news.tsv chứa thông tin các tin tức

3. Tệp entity\_embedding.vec chứa các vec-tơ nhúng (embedding) của các thực thể trong tin tức, dựa trên Wikidata

4. relation\_embedding.vec chứa các vec-tơ nhúng của mối quan hệ giữa các thực thể Được sử dụng phổ biến trong các mô hình gợi ý tin tức là hai tệp behaviors.tsv và news.tsv.

Bảng dưới mô tả chi tiết các trường thuộc tính trong tệp behaviors.tsv. Lưu ý rằng, trong tệp không có phần tiêu đề, và thứ tự các trường thuộc tính dưới đây khớp với thứ tự thuộc tính trong tệp.

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Thuộc tính | Ý nghĩa thuộc tính | Ví dụ minh họa |
| 1 | ImpreessionID | ID của một hiển thị cho người dùng có mã là UserID | 1 |
| 2 | User ID | ID (ẩn danh) của người dùng | U134050 |
| 3 | Time | Thời gian hiển thị với định dạng "MM/DD/YYYY HH:MM:SS AM/PM" | 11/15/2019 8:55:22 AM |
| 4 | History | Lịch sử nhấp chuột trên tin tức (danh sách các ID của tin tức đã được nhấp) của người dùng trước hiển thị này (được sắp xếp theo thời gian) | N12246 N128820 N119226 N4065 N67770 N33446 N103285 N99640 N106837 N76775 N129444 N120026 N115495 N86141 N98680 N108013 N121782 N30276 N93576 N123743 N48804 N79909 N68728 N71356 |
| 5 | Impression | Danh sách tin tức hiển thị trong impression và hành vi nhấp chuột của người dùng (1-click, 0-no click) (đã được xáo trộn) | N91737-0 N30206-0 N54368-0 N117802-0 N18190-0 N122944-0 N69938-1 N18356-0 N123209-0 N46894-0 N29160-0 N49978-0 N89764-0 N30582-0 N58465-0 N35304-0 |

Bảng dưới mô tả chi tiết các trường thuộc tính trong tệp news.tsv. Lưu ý rằng, trong tệp không có phần tiêu đề, và thứ tự các trường thuộc tính dưới đây khớp với thứ tự thuộc tính trong tệp.

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Thuộc tính | Ý nghĩa thuộc tính | Ví dụ minh họa |
| 1 | News ID | ID của tin tức | N1387 |
| 2 | Category | Danh mục của tin tức | entertainment |
| 3 | SubCategory | Danh mục con của tin tức | gaming |
| 4 | Title | Tiêu đề của tin tức | Best PS4 games 2019: play great PlayStation 4 games |
| 5 | Abstract | Phần tóm tắt hoặc trích dẫn của tin tức | These are the best PS4 games available today, from Sony's awesome exclusives to third-party must plays |
| 6 | URL | Địa chỉ URL của tin tức trên mạng | https://assets.msn.com/labs/mind/AACkFyc.html |
| 7 | Title Entities | Các thực thể (entities) có trong tiêu đề của tin tức | [{"Label": "PlayStation 4", "Type": "J", "WikidataId": "Q5014725", "Confidence": 1.0, "OccurrenceOffsets": [5, 32], "SurfaceForms": ["PS4", "PlayStation 4"]}] |
| 8 | Abstract Entities | Các thực thể (entities) có trong phần tóm tắt của tin tức | [{"Label": "PlayStation 4", "Type": "J", "WikidataId": "Q5014725", "Confidence": 1.0, "OccurrenceOffsets": [19], "SurfaceForms": ["PS4"]}, {"Label": "Sony", "Type": "O", "WikidataId": "Q41187", "Confidence": 0.936, "OccurrenceOffsets": [51], "SurfaceForms": ["Sony"]}] |

Một số thống kê về tập dữ liệu này:

Phân tích File news.tsv (MIND-large train):

Định dạng thông tin về các bài báo:

Ảnh có chứa văn bản

Mô tả được tạo tự động

Thống kê qua về file:

Ảnh có chứa bàn

Mô tả được tạo tự động

Từ hình trên cho thấy:

Có tất cả 101527 bài báo

Thông tin các trường đa số đủ, trong đó:

Trường Abstract có 5412 hàng bị NaN

Trường Title Entities có 3, trường Abstract Entities có 6 hàng bị NaN.

Phân tích về danh mục (category) và danh mục con (subcategory):

A graph of blue squares

Description automatically generated

Trong 101527 bài báo có :

32020 bài về thể thao (sports)

30478 bài về tin tức thời sự (news)

5916 bài về tài chính

4955 bài về du lịch

4569 bài về phong cách sống

….

Từ đó cho thấy thể thao và tin tức là 2 danh mục được đa số mọi người quan tâm.

Bên cạnh các danh mục (category) như đã thống kê trên thì trong mỗi danh mục còn có nhiều danh mục con (subcategory):

A graph of a graph

Description automatically generated with medium confidence

Như hình trên có thể thấy trong loạt báo viết về thể thao thì báo được quan tâm nhiều nhất là báo về bóng đá, tiếp theo là bóng chày, bóng rổ,… theo đúng như độ yêu thích của các môn thể thao này trong thực tế.

Còn trong những bài báo về thời sự (danh mục báo được quan tâm nhiều nhất) thì các bài viết về thời sự xung quanh chúng ta (newsus), sau đó là các tin tức về chính trị, khoa học,…

Danh mục thời tiết thì không có danh mục con vì chỉ có thông tin về thời tiết hàng ngày.

Danh mục video thì các bài báo video về thời sự cũng được yêu thích hàng đầu, tiếp theo là động vật, khoa học,…

Các bài báo về lĩnh vực trẻ em, games, giải trí, phim, nhạc,TV,.. rất ít được quan tâm vì theo lẽ tự nhiên tin tức về trẻ em thì chỉ thu hút các bà mẹ mang bầu (có lẽ), games, phim, nhạc, TV, giải trí thì người ta sẽ thực hành luôn trên các trang chuyên về các lĩnh vực thay vì đọc báo về chúng.

Đặc biệt các báo mang danh mục vùng miền, quốc gia như middle\_east, north\_america thì cũng cực ít người quan tâm (1 người/danh mục)

Phân tích về title của 2 danh mục được đông đảo người dùng quan tâm nhất:

Title là cái chính nhất thu hút người dùng có đọc bài báo hay không vì thế nó đóng vai trò vô cùng quan trọng. Do vậy, các tác giả phải chọn lọc các từ sao cho nói lên ý chính của bài báo để thông báo cho người dùng biết bài báo này sẽ nói về vấn đề gì.

WordCloud của các title trong danh mục Thể thao (sports) sau khi đã lược bỏ đi từ dừng và chuyển về lemma:

Ảnh có chứa văn bản

Mô tả được tạo tự động

Như wordcloud về title trên có thể thấy các từ quan trọng được nổi bật như “win”, ”loss’, “game”, “team”, “week”,.. được dùng nhiều để làm title của các bài báo thể thao.WordCloud của các title trong danh mục tin tức thời sự:

Ảnh có chứa văn bản

Mô tả được tạo tự động

Nổi bật trên các từ được dùng để đặt cho title của bài báo về tin tức thời sự là “Trump”. Có thể nhận thấy Trump là tổng thống thứ 45 của Hoa Kỳ và điều đặc biệt năm 2019 (năm lấy mẫu dữ liệu) là một năm đầy sóng gió của tổng thống Donal Trump.

Độ dài của các title:

A blue line graph

Description automatically generated

Độ dài của các title có dáng điệu giống như phân phối chuẩn, phản ảnh độ dài title không nên quá dài hay quá ngắn.

Phân tích file behaviors.tsv:

File có 5 trường được mô tả trên phần tổng quan:

Ảnh có chứa văn bản

Mô tả được tạo tự động

File gồm 2232748 bản ghi, ghi lại lịch sử click của người dùng, trong đó trường History (lịch sử click của người dùng trước thời điểm ghi logs) có 46065 bản ghi, đây là những người chưa đọc bài báo nào trước đó (người dùng mới).

A screenshot of a computer code

Description automatically generated

Thống kê số lượng positive và negative:

A graph with blue bars

Description automatically generated with medium confidence

Trong đó:

Positvie : 338656

Negative : 80123718

Ratio : 0.04223039175491083

Thống kê users:

Tập train:

A blue circle with text on it

Description automatically generated

Số lượng normal user chiếm phần lớn với tỉ lệ 98.2% tương ứng với 711222 user.

Tập test:

A pie chart with text and numbers

Description automatically generated

Cold start user ở tập test chiếm nhiểu hơn trong tập train với 4.1% tương ứng với 29108 user, do đó mô hình dự toán nếu dự đoán được các cold start user này thích bài báo nào hơn bài báo nào thì score sẽ được cải thiện.

* 1. **Tiền xử lý dữ liệu**

Để thuận tiện trong quá trình huấn luyện, em thực hiện một số hàm tiền xử lý dữ liệu sau:

**Parse\_behaviors:**

Chuyển đổi trường chỉ mang tính chất định danh như userID thành int và lưu ánh xạ vào 1 file user2int.tsv

Loại bỏ trường time.

Thêm 1 trường clicked và bỏ clicked ở trường candidates news cũ.

VD: N91737-0 N30206-0 được tách thành 2 trường candidates news: N91737 N30206 và clicked: 0 0

**Parse\_news:**

Trường category và subcategory được chuyển về kiểu int dưới 1 file ánh xạ category2int.

Xây dựng từ điển word2int và chuyển title, abstract theo từ điển.

**Generate\_words\_embedding:**

Tạo ra ma trận word embedding từ pretrained Glove và từ điển word2int ở trên.

1. **CÁC PHƯƠNG PHÁP ĐÁNH GIÁ**
   1. **Các tham số đánh giá**

Để đánh giá hiệu quả của mô hình, em sử dụng 4 độ đo đánh giá bao gồm: AUC, MRR, nDCG@5 và nDCG@10.

**Chỉ số AUC (Area Under the Curve)**

Bắt đầu với đường cong ROC. ROC là viết tắt của Receiver Operating Characteristic, là một biểu diễn đồ họa về hiệu suất của mô hình phân loại nhị phân, thể hiện tỷ lệ dương tính thực sự (True Positive Rate, TPR) so với tỷ lệ dương tính giả (False Positive Rate, FPR) ở các cài đặt ngưỡng khác nhau.

TPR là tỷ lệ phần trăm của các mẫu dương được dự đoán chính xác trong số tất cả các mẫu dương thực tế, được tính bởi công thức:



FPR là tỷ lệ phần trăm của các mẫu dương được dự đoán không chính xác trong số tất cả các mẫu âm thực tế, được tính bởi công thức:



Ngưỡng là giá trị trên đó dự đoán trở nên dương hoặc âm tùy thuộc vào thiết lập, ví dụ ngưỡng không nhỏ hơn 0.5, tức là mọi mẫu đạt trên 50% cho lớp đích sẽ trở thành lớp này. Một mô hình tốt sẽ có một đường cong ROC ôm lấy góc trên cùng bên trái của biểu đồ, có nghĩa là có TPR cao và FPR thấp ở tất cả các cài đặt ngưỡng. Một mô hình đưa ra các dự đoán ngẫu nhiên sẽ có đường cong ROC là một đường chéo từ góc dưới cùng bên trái đến góc trên cùng bên phải của biểu đồ, có nghĩa là TPR và FPR của nó sẽ bằng nhau ở tất cả các cài đặt ngưỡng.

A line graph with a point

Description automatically generated with medium confidence

Chỉ số AUC được tính bằng cách tìm diện tích dưới đường cong của đồ thị ROC. Một mô hình hoàn hảo sẽ có AUC bằng 1, có nghĩa là tất cả các mẫu dương được xếp hạng cao hơn so với tất cả các mẫu âm, trong khi một mô hình dự đoán ngẫu nhiên sẽ có AUC bằng 0.5. Điều này là quan trọng vì nó làm cho AUC trở thành một thước đo tốt hơn so với chỉ số Accuracy đơn giản, thứ chỉ là tỷ lệ phần trăm của dự đoán đúng so với tất cả các đầu vào có thể có trong tập dữ liệu. Hình sau mô tả cụ thể phần diện tích cần tính (tô màu) là chỉ số AUC.

A blue and orange line graph

Description automatically generated

**Chỉ số xếp hạng đối ứng trung bình MRR (Mean Reciprocal Rank)**

MRR đánh giá danh sách các mục được xếp hạng cho các truy vấn. Thứ hạng đối ứng là "nghịch đảo nhân" của thứ hạng của mục chính xác đầu tiên. Để tính toán MRR, các mục không cần phải được xếp hạng. MRR không áp dụng nếu có nhiều phản hồi đúng (lần truy cập) trong danh sách kết quả. Chỉ số MRR được tính bởi công thức:



trong đó, 𝑟𝑎𝑛𝑘௜ là thứ hạng của phản hồi đúng trong |𝑄| truy vấn.

**Chỉ số nDCG@5 và nDCG@10**

𝑛𝐷𝐶𝐺 (Normalized Discounted Cumulative Gain) được thiết kế để đo lường hiệu quả của các mô hình xếp hạng bằng cách đánh giá chất lượng của danh sách kết quả hoặc dự đoán được sắp xếp theo thứ tự. Cả mức độ liên quan của từng kết quả và vị trí của nó trong danh sách đều góp phần trong tính toán 𝑛𝐷𝐶𝐺.

𝐷𝐶𝐺 (Discounted Cumulative Gain) là nền tảng của 𝑛𝐷𝐶𝐺, kết hợp sự liên quan của kết quả truy vấn và vị trí của nó trong danh sách kết quả. 𝐷𝐶𝐺 ở một thứ hạng cụ thể 𝑝 được tính bởi công thức:



trong đó,  là điểm phù hợp được đánh giá cho mục 𝑖 dựa trên mức độ liên quan của nó trong tìm kiếm (thường là 0 nếu không liên quan, 1 nếu liên quan).

Thực tế, khi người dùng tìm kiếm thứ gì đó trực tuyến, họ luôn muốn có mục phù hợp nhất ở trên cùng và phía trên bất kỳ mục không liên quan nào. Đó là, tất cả các thông tin liên quan phải luôn ở trên cùng và nó phải có 𝐷𝐶𝐺 tốt nhất, do đó, ta có chỉ số 𝐼𝐷𝐶𝐺 (Ideal Discounted Cumulative Gain), được tính bởi công thức:



trong đó, 𝑅 là trạng thái gợi ý lý tưởng (các mục liên quan đều được ở trên đầu danh sách), 𝑅௜ là điểm phù hợp được đánh giá cho mục 𝑖 dựa trên mức độ liên quan của nó trong trạng thái lý tưởng (tốt nhất).

Chỉ số chuẩn hóa về khoảng từ 0 đến 1, do đó:



Nếu tất cả các mục được xếp hạng hoàn hảo, bằng 1. Nó cung cấp một cách để đo lường hiệu suất của hệ thống đề xuất hoặc bảng xếp hạng kết quả tìm kiếm, đặc biệt tập trung vào 𝑝 vị trí hàng đầu. Đối với chỉ số, ta cho 𝑝 = 5. Đối với chỉ số, ta cho 𝑝 = 10.

* 1. **Môi trường thực nghiệm**

Em thực hiện xây dựng và huấn luyện mô hình trên môi trường Kaggle. Một số thông số về cấu hình trong môi trường chạy:

• GPU: Sử dụng GPU T4 , bộ nhớ 16 GB.

• CPU: RAM 29 GB

* 1. **Kết quả huấn luyện**

Em lựa chọn các tham số chung cho các mô hình như sau:

LR = 1e-4

Opt = Adam

Loss : Cross Entropy

Title = 20

Abstract = 100

Epoch = 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình** | **AUC** | **MRR** | **nDCG@5** | **nDCG@10** |
| NAML | **0,6854** | **0,3372** | **0,3672** | **0,4243** |
| NRMS | 0,6637 | 0,3123 | 0,3477 | 0,4045 |

1. **KẾT LUẬN**

Thông qua bài tập lớn này, em đã thực hiện thử nghiệm một số mô hình về gợi ý tin tức, tìm hiểu mô hình gợi ý dựa trên học đa tác vụ và cài đặt mô hình dựa trên bài báo gốc. Tổng kết lại, em đã nắm được luồng xử lý của một bài toán hệ gợi ý, nắm được cách xử lý dữ liệu cho các mô hình khác nhau, thử nghiệm tính hiệu quả của việc học đa tác vụ cho bài toán gợi ý tin tức. Kết quả thu được tuy chưa vượt được kỳ vọng, nhưng cũng đã phản ánh được khả năng của phương pháp tiếp cận này.

Vấn đề khó khăn nhất mà em gặp phải là giới hạn thời gian và tài nguyên. Cụ thể, môi trường Kaggle chỉ cho phép chạy 12 tiếng liên tục cho một phiên, cùng với giới hạn bộ nhớ GPU ở mức 16 GB khiến một số mô hình em thử nghiệm, ví dụ thêm đặc trưng, tinh chỉnh siêu tham số, kết hợp mô hình trở nên không khả thi.