ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO BÀI TẬP CUỐI KỲ
KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: PHÂN LOẠI VÀ ĐÁNH NHÃN MÃ BỆNH DỰA TRÊN NHỮNG GHI CHÚ BỆNH ÁN

Ngày 14 tháng 6 năm 2022

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO BÀI TẬP CUỐI KỲ

Tin sinh học

ĐỀ TÀI: PHÂN LOẠI VÀ ĐÁNH NHÃN MÃ BỆNH DỰA TRÊN NHỮNG GHI CHÚ BỆNH ÁN

Giảng viên: TS Đặng Cao Cường

Sinh viên: Trần Công Việt An - 19020032
Nguyễn Thế Quân - 18021030
Nguyễn Quang Huy - 16020999

Ngày 14 tháng 6 năm 2022

Lời cam đoan

Chúng em cam đoan rằng bản báo cáo này hoàn toàn được hoàn thành dựa trên sự cố gắng, nỗ lực của tất cả các thành viên trong nhóm.

Báo cáo này hoàn toàn không sao chép từ bất kỳ một nguồn nào khác. Nếu phát hiện có sự gian lận, chúng em xin chịu hoàn toàn trách nhiệm.

Trần Công Việt An **Nguyễn Thế Quân**

Nguyễn Quang Huy

Danh sách thành viên

MSV	Họ và Tên	Phân công công việc
19020032	Trần Công Việt An	• Nhóm trưởng • Phân tích kiến trúc mô hình
18021030	Nguyễn Thế Quân	• Phân tích bài toán
16020999	Nguyễn Quang Huy	• Phân tích kết quả đánh giá mô hình

Mục lục

Danh sách hình vẽ

Tóm tắt nội dung

Ghi chú y tế là các loại văn bản được tạo ra bởi các nhân viên y tế, được thực hiện cho mỗi lần gặp bệnh nhân. Chúng thường đi kèm với mã y tế, những mã số mô tả các chẩn đoán và cách điều trị. Việc đánh nhãn mã y tế cho các văn bản này rất tốn công sức và dễ gặp lỗi; hơn nữa, mối liên kết giữa văn bản và mã thường không được ghi lại rõ ràng, bỏ qua lý do và chi tiết đằng sau chẩn đoán và điều trị cụ thể. Chúng tôi xin giới thiệu một mạng học sâu với cơ chế chú ý (attention) và tích chập (convolutional) có thể dự đoán được các mã y tế trong văn bản ghi chú y tế. Phương pháp của chúng tôi tổng hợp thông tin trên toàn bộ một văn bảng bằng cách sử dụng một mạng nơ-ron tích chập, và sử dụng cơ chế chú ý để chọn các phân đoạn văn bản có liên quan nhất cho mỗi một trong số hàng ngàn nhãn. Phương pháp này cho độ chính xác cao, đạt được độ chính xác@8 là 0,71 và Micro-F1 là 0,54, cả hai đều tốt hơn mô hình tốt nhất trước đó. Hơn nữa, thông qua đánh giá khả năng diễn giải bởi một bác sĩ, chúng tôi cho thấy rằng việc sử dụng cơ chế chú ý có thể phát hiện ra những lời diễn giải hợp lý cho mỗi mã y tế được gắn.

1 Giới thiệu

Ghi chú lâm sàng là các văn bản mang tính chất tường thuật tự do được viết bởi các bác sĩ lâm sàng trong các buổi khám bệnh với bệnh nhân. Chúng thường đi kèm với một bộ mã từ Bảng phân loại bệnh tật quốc tế (ICD), là một phương pháp tiêu chuẩn để xác định các chẩn đoán và quy trình đã được thực hiện trong cuộc buổi khám bệnh. Mã ICD có nhiều mục đích sử dụng, từ hỗ trợ tính toán viện phí đến sử dụng làm dữ liệu đầu vào cho mô hình dự đoán trạng thái bệnh nhân (Choi và cộng sự, 2016; Ranganath và cộng sự, 2015; Denny và cộng sự, 2010; Avati và cộng sự, 2017). Bởi vì đánh mã thủ công là một quá trình tốn thời gian và dễ xảy ra lỗi, mã hóa tự động đã được nghiên cứu ít nhất từ những năm 1990 (de Lima và cộng sự, 1998).

Đây là một bài toán khó giải quyết vì hai lý do chính. Đầu tiên, không gian nhãn có chiều rất cao, với hơn 15.000 mã trong phân loại ICD-9 và hơn 140.000 mã tổng cộng trong phiên bản mới hơn ICD-10-CM và ICD-10-PCS (Tổ chức Y tế Thế giới, 2016). Thứ hai, ghi chú y tế thường mang cả thông tin không liên quan, lỗi chính tả và các chữ viết tắt không chuẩn, và thường sử dụng một lượng lớn từ vựng chuyên ngành. Các đặc điểm này này kết hợp với nhau khiến cho dự đoán mã ICD từ các ghi chú lâm sàng là một bài toán đặc biệt khó khăn, kể cả đối với máy tính và con người (Birman-Deych và cộng sự, 2005). Trong bài viết ứng dụng này, chúng tôi phát triển mô hình dựa trên mạng nơ-ron tích chập (CNN) cho việc gán mã ICD tự động dựa trên phân loại văn bản xuất viện từ đơn vị chăm sóc đặc biệt (ICU). Để thích ứng tốt hơn với bài toán nhiều nhãn, chúng tôi sử dụng cơ chế chú ý theo từng nhãn, cho phép mô hình của chúng tôi tìm hiểu học được cách biểu diễn văn bản riêng biệt cho mỗi nhãn khác nhau.

Chúng tôi gọi phương pháp này là Convolutional Attention for Multi-Label classification (CAML). Thiết kế mô hình của chúng tôi được thúc đẩy dựa trên phỏng đoán rằng các thông tin quan trọng liên quan đến sự hiện diện của mã có thể được chứa trong các đoạn văn bản ngắn, ở bất kỳ đâu trong tài liệu, và những đoạn mã này có thể khác nhau cho các nhãn khác nhau. Để đối phó với không gian nhãn lớn, chúng tôi khai thác các mô tả văn bản của mỗi mã để hướng dẫn mô hình của chúng tôi điều chỉnh tham số: trong trường hợp không có nhiều ví dụ được gắn nhãn đối với một mã nhất định, các tham số của nó phải tương tự đối với những mã có mô tả tương tự. Chúng tôi đánh giá cách tiếp cận của mình trên hai phiên bản của MIMIC (Johnson và cộng sự, 2016), một tập dữ liệu y tế công cộng. Mỗi bản ghi bao gồm một loạt các ghi chú tường thuật mô tả một bệnh nhân, bao gồm cả các chẩn đoán và thủ tục y tế được thực hiện. Hướng tiếp cận của chúng tôi về cơ bản làm tốt hơn rất nhiều so với các kết quả trước đó khi dự đoán mã y tế trên cả MIMIC-II và MIMIC-III.

Chúng tôi xem xét các ứng dụng của mô hình này trong một bối cảnh hỗ trợ đưa ra quyết định. Khả năng diễn giải kết quả là rất quan trọng cho bất kỳ hệ thống hỗ trợ quyết định nào, đặc biệt là trong lĩnh vực y tế. Hệ thống sẽ cần phải giải thích rõ ràng tại sao nó dự đoán từng mã; ngay cả khi các mã được đánh nhãn theo cách thủ công, bạn nên giải thích những phần nào của văn bản có liên quan nhất đến từng mã số. Những yêu cầu này tiếp tục thúc đẩy cơ chế chú ý trên từng nhãn. Cơ chế này đưa ra mức độ quan trọng cho từng n-gram trong tài liệu đầu vào, và do đó có thể cung cấp giải thích cho mỗi mã, dưới dạng các đoạn văn bản được trích xuất từ tài liệu đầu vào. Chúng tôi thực hiện đánh giá chất lượng bởi một bác sĩ cho các cách diễn giải được hỗ trợ bởi cơ chế chú ý, yêu cầu bác sĩ đó đánh giá mức độ hàm chứa thông tin của các diễn giải này.

2 Phương pháp

Chúng tôi coi bài toán dự đoán mã ICD-9 như một bài toán phân loại văn bản đa nhãn (McCallum, 1999). Cho $\mathcal{L}$ biểu diễn toàn bộ bộ mã ICD-9; bài toán đánh nhãn cho một văn bản i cần xác định, $y_i, l \in 0, 1$ với $l \in \mathcal{L}$. Chúng tôi huấn luyện một mạng nơ-ron chuyển văn bản qua một lớp tích chập để học được một đại diện cơ sở cho mỗi văn bản (Kim, 2014) và tạo ra $\mathcal{L}$ quyết định phân loại nhị phân. Thay vì tổng hợp trên đại diện này với một lớp max-polling, chúng tôi áp dụng cơ chế chú ý (attention mechanism) để chọn các phần của tài liệu có liên quan nhất cho mỗi mã. Các trọng số chú ý này sau đó được áp dụng vào các biểu diễn cơ sở và kết quả được đưa qua một lớp đầu ra, sử dụng hàm kích hoạt sigmoid để tính toán khả năng xuất hiện của từng mã. Chúng tôi sử dụng một hàm chính quy hóa (regularizer) để khuyến khích tham số của các mã có mô tả tương tự nhau trở nên giống nhau. Chúng tôi sẽ mô tả những yếu tố này chi tiết hơn trong phần còn lại của bài báo.

2.1 Kiến trúc hàm tích chập

Trong lớp gốc của mô hình, chúng ta có một lớp embedding d_e chiều được huấn luyện cho mỗi từ ở trong văn bản. Chúng được biểu diễn nối lại thành một ma trân $\boldsymbol{X} = [x_1, x_2, \ldots, x_N]$, trong đó N là chiều dài của văn bản. Các mã hóa word embeddings kề nhau được gộp lại sử dụng một lớp lọc tích chập $\boldsymbol{W}_c \in \mathbb{R}^{k \times d_e \times d_c}$, trong đó k là chiều rộng của lớp lọc, d_e là kích cỡ của lớp embedding, và d_c là kích cỡ đầu ra của lớp tích chập. Ở mỗi bước n, ta tính

$$\boldsymbol{h}_n = g\left(\boldsymbol{W}_c * \boldsymbol{x}_{n:n+k-1} + \boldsymbol{b}_c\right)$$

trong đó $*$ biểu diễn thao tác nhân tích chập, g là một hàm kích hoạt phi tuyến tính, và $\boldsymbol{b}_c \in \mathbb{R}^{d_c}$ là tham số bias. Chúng tôi cũng đệm thêm số 0 ở các cạnh bên của đầu vào để cho $\boldsymbol{H}$ có kích cỡ $\mathbb{R}^{d_c \times N}$.

2.2 Lớp chú ý

Sau lớp tích chập, văn bản hiện đang được biểu diễn bởi ma trận $\boldsymbol{H} \in \mathbb{R}^{d_c \times N}$. Thông thường, ta sẽ biến toàn bộ ma trận này thành một vector bằng cách áp dụng pooling xuyên suốt chiều dài của văn bản, bằng cách lấy giá trị lớn nhất (max-pool) hoặc trung bình (average-pool) ở mỗi hàng. Tuy nhiên mục tiêu của chúng tôi là gán nhiều nhãn cho mỗi văn bản, và mỗi phần khác nhau của biểu diễn văn bản gốc có thể liên quan đến những nhãn khác nhau. Vì lý do này, chúng tôi áp dụng cơ chế chú ý dựa trên từng nhãn. Một lợi thế phụ là nó cũng sẽ chọn ra các đoạn k-grams từ văn bản mà liên quan nhất đến các nhãn đang được dự đoán.

Mô tả bài toán một cách hình thức, đối với mỗi nhãn ℓ, chúng tôi tính $\boldsymbol{H}^\top \boldsymbol{u}_\ell$, trong đó $\boldsymbol{u}_\ell \in \mathbb{R}^{d_c}$ là một tham số vector dành cho nhãn ℓ. Sau đó, chúng tôi đưa vector kết quả qua một hàm softmax, thu được một phân bố xác suất ở các vị trí trong văn bản.

$$\boldsymbol{\alpha}_\ell = \text{SoftMax}\left(\boldsymbol{H}^\top \boldsymbol{u}_\ell\right),$$

trong đó $\text{SoftMax}(x) = \frac{\exp(x)}{\sum_i \exp(x_i)}$, và $\exp(\boldsymbol{x})$ là hàm mũ tự nhiên từng thành phần cho $\boldsymbol{x}$. Vector chú ý $\boldsymbol{\alpha}$ sau đó được sử dụng để tính toán biểu diễn vector cho mỗi nhãn,

$$\boldsymbol{v}_\ell = \sum_{n=1}^{N} \alpha_{\ell,n} \boldsymbol{h}_n$$

Để có một mô hình cơ sở, chúng tôi dùng max-pooling để tính toán một vector v cho mọi nhãn,

$$v_j = \max_n h_{n,j}$$

2.3 Phân loại

Cho biểu diễn vector của văn bản $\boldsymbol{v}_\ell$, chúng tôi tính toán ra xác suất cho nhãn ℓ sử dụng một lớp tuyến tính và một hàm kích hoạt sigmoid.

$$\hat{y}_\ell = \sigma\left(\boldsymbol{\beta}_\ell^\top \boldsymbol{v}_\ell + b_\ell\right),$$

trong đó $\boldsymbol{\beta}_\ell \in \mathbb{R}^{d_c}$ là một vector các trọng số dự đoán, and b_ℓ là một giá trị bias vô hướng. Toàn bộ kiến trúc được thể hiện trong 2.1

2.4 Huấn luyện mô hình

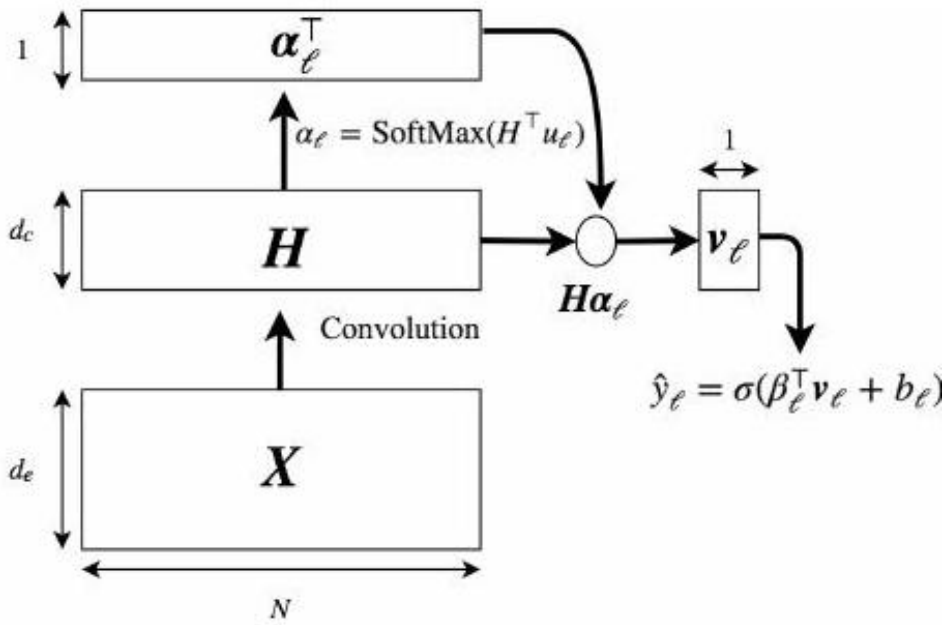
Quy trình huấn luyện tối ưu hàm mục tiêu cross-entropy nhị phân,

$$L_{\text{BCE}}(\boldsymbol{X}, \boldsymbol{y}) = -\sum_{\ell=1}^{\mathcal{L}} y_\ell \log\left(\hat{y}_\ell\right) + (1 - y_\ell) \log\left(1 - \hat{y}_\ell\right),$$

cộng với hàm trung bình L2 của tham số mô hình.

2.5 Nhúng biểu diễn của nhãn

Do kích thước của không gian nhãn, nhiều mã hiếm khi được quan sát trong dữ liệu được gán nhãn. Để cải thiện hiệu năng của mô hình trên các mã này, chúng tôi sử dụng mô tả của từng mã trong từ điển của Tổ chức Y tế Thế giới (2016). Ví dụ có thể được tìm



Hình 2.1: Kiến trúc CAML cho một nhãn. Trong một kiến trúc max-pooling, $\boldsymbol{H}$ được ánh xạ thẳng đến vector $\boldsymbol{v}_\ell$ bằng cách lấy giá trị tối đa trên từng chiều

thấy trong Bảng 1, bên cạnh các mã số. Chúng tôi sử dụng những mô tả đó để xây dựng một module phụ trong mạng học sâu. Module này học cách nhúng các mô tả đó dưới dạng vectơ. Các vectơ này sau đó được sử dụng làm mục tiêu của hàm chuẩn hóa trên các tham số của mô hình $\boldsymbol{\beta}_\ell$. Nếu mã ℓ hiếm khi được quan sát trong dữ liệu huấn luyện, hàm chuẩn hóa sẽ khuyến khích các thông số của nó trở nên tương tự như những mã khác có mô tả tương tự.

Module nhúng mã bao gồm một lớp CNN max-pooling. Cho $\boldsymbol{\beta}_\ell$ là một vector được max-pool, có được bằng cách chuyển mô tả của nhãn ℓ vào module nhúng mã. Cho n_y là số nhãn đúng trong một ví dụ huấn luyện. Chúng tôi đưa vào hàm mục tiêu của việc huấn luyện một thành phần chuẩn hóa như sau:

$$L(\boldsymbol{X}, \boldsymbol{y}) = L_{\text{BCE}} + \lambda \frac{1}{n_y} \sum_{\ell: y_\ell = 1}^{\mathcal{L}} \|\boldsymbol{z}_\ell - \boldsymbol{\beta}_\ell\|_2 ,$$

trong đó λ là một siêu tham số để cân đối ảnh hưởng của hai thành phần mục tiêu. Chúng tôi gọi biến thể này là Description Regularized CAML (DR-CAML)

3 Đánh giá kết quả của phương án tiếp cận

Thực hiện đánh giá độ chính xác của mô hình đề xuất bởi tác giả, và so sánh với một số mô hình cơ sở.

3.1 Tập dữ liệu đánh giá

Tập dữ liệu MIMIC-III sẽ được sử dụng để huấn luyện và đánh giá. Trong tập dữ liệu này, tác giả sẽ tập trung vào các đoạn văn bản ghi chú bệnh án của người bệnh.

Dữ liệu mỗi lần nhập viện đều được người mã hóa gắn thẻ với một bộ mã ICD-9, mô tả cả những chẩn đoán và các quy trình được thực hiện trong thời gian bệnh nhân nằm viện. Có 8.921 mã ICD-9 xuất hiện trong bộ dữ liệu, bao gồm 6.918 mã chẩn đoán và 2.003 mã quy trình. Một số bệnh nhân có nhiều lần nhập viện và do đó có nhiều bản ghi tổng hợp khi xuất viện; tác giả đã chia dữ liệu theo định danh bệnh nhân để không có bệnh nhân nào xuất hiện trong cả tập huấn luyện và tập dữ liệu thử nghiệm.

Tác giả đã chia tập dữ liệu ra thành các bộ dữ liệu huấn luyện, kiểm thử. Tập dữ liệu đào tạo bao gồm 47.724 bản ghi tổng hợp từ 36.998 bệnh nhân. Và hai bộ gồm 1.632 bản ghi và 3.372 bản ghi sử dụng để thẩm định và kiểm thử.

Ngoài ra, để so sánh kết quả với những mô hình đã tồn tại, tác giả có tham chiếu thêm các mô hình của Shi **andothers** 2017, Baumel **andothers** 2017. Nhưng do các mô hình được huấn luyện trên những tập dữ liệu khác nhau, như mô hình của Shi **andothers** 2017 được huấn luyện trên tập 50 mã phổ thông nhất của tập dữ liệu MIMIC-III hay mô hình của Baumel **andothers** 2017 được huấn luyện trên tập dữ liệu MIMIC-II, nên tác giả đã chạy trên cả những tập dữ liệu trên để có những kết quả so sánh chính xác nhất.

Tiền xử lý Tác giả đã xóa các ký tự không chứa ký tự chữ cái (ví dụ: xóa "500" nhưng vẫn giữ "250mg"), chuyển tất cả các mã thông báo thành chữ thường và thay thế các ký tự xuất hiện trong ít hơn ba tài liệu đào tạo bằng ký tự 'UNK'. Chúng tôi chuẩn bị trước từ nhúng có kích thước $d_e = 100$ bằng phương pháp CBOW word2vec trên văn bản được xử lý trước từ tất cả các bản báo cáo tổng hợp xuất viện. Tất cả các tài liệu được cắt ngắn đến độ dài tối đa là 2500 ký tự.

3.2 System

Tác giả so sánh mô hình với các baseline sau:

- Một mạng tích chập đơn (Kim, 2014)
- Một mô hình hồi quy tuyến tính với túi từ vựng (bag of words)
- Một mạng Bi-GRU

Đối với CNN và Bi-GRU, chúng tôi khởi tạo trọng số lớp nhúng bằng cách sử dụng cùng một mạng nhúng word2vec được đào tạo trước cho mô hình CAML. Tất cả các các mô hình được xây dựng bằng thư viện PyTorch. Mô hình hồi quy logistic bao gồm $|\mathcal{L}|$ bộ phân loại một-vs-còn lại hoạt động trên các túi từ vựng unigram cho tất cả các nhãn có trong tập dữ liệu huấn luyện. Nếu một nhãn không có trong tập dữ liệu huấn luyện, mô hình sẽ không bao giờ dự đoán ra nó trong các dữ liệu được đưa ra sau này.

Tinh chỉnh tham số Chúng tôi điều chỉnh siêu tham số cho mô hình CAML và các mô hình cơ sở bằng cách sử dụng thư viện tối ưu hóa Spearmint Bayesian (Snoek và cộng sự, 2012; Swersky và cộng sự, 2013). Chúng tôi cho phép Spearmint lấy mẫu các giá trị tham số cho hàm mất mát L2 trên trọng số của mô hình ρ và tốc độ huấn luyện η, cũng như kích thước bộ lọc k, số lượng bộ lọc d_c và xác suất dropout q cho mô hình tích chập và số lớp ẩn s có số chiều v của mô hình Bi-GRU, sử dụng precision@8 trên tập dữ liệu xác thực được trích xuất ra từ MIMIC-III để làm độ đo hiệu năng. Chúng tôi sử dụng các thông số này cho DR-CAML cũng như chuyển các thông số được tối ưu hóa sang các mô hình cho MIMIC-II đầy đủ nhãn và MIMIC-III 50 nhãn và tinh chỉnh thủ công tốc độ huấn luyện trong các cài đặt này. Chúng tôi chọn λ cho DR-CAML dựa trên thử nghiệm thí điểm trên tập dữ liệu xác thực được tóm tắt trong 4. Mô hình tích chập được huấn luyện với kỹ thuật dropout sau lớp embedding. Chúng tôi sử dụng kích cỡ batch là 16 cho tất cả các mô hình. Mô hình được huấn luyện với kỹ thuật early stopping: quy trình huấn luyện sẽ kết thúc sau 10 chu kỳ không tối ưu được thêm tham số precision@8, và mô hình có precision@8 cao nhất sẽ được sử dụng làm mô hình kết quả.

3.3 Phương thức đánh giá

Để dễ dàng so sánh với cả công việc trong tương lai và trước đây, chúng tôi đánh giá dựa trên nhiều số liệu khác nhau, tập trung vào micro-average F1 và marco-average F1 và khu vực dưới đường cong ROC (AUC). Giá trị micro-average F1 được tính bằng cách coi mỗi cặp (văn bản, mã) là một dự đoán riêng biệt. Giá trị marco-average F1, mặc dù ít được báo cáo hơn trong tài liệu phân loại đa nhãn, nhưng được tính bằng cách tính số liệu trung bình cho mỗi nhãn. Để đưa ra kết quả đánh giá, các chỉ số được phân biệt như sau:

$$\text{Micro-R} = \frac{\sum_{\ell=1}^{|\mathcal{L}|} \text{TP}_\ell}{\sum_{\ell=1}^{|\mathcal{L}|} \text{TP}_\ell + \text{FN}_\ell}$$

$$\text{Macro-R} = \frac{1}{|\mathcal{L}|} \sum_{\ell=1}^{|\mathcal{L}|} \frac{\text{TP}_\ell}{\text{TP}_\ell + \text{FN}_\ell},$$

trong đó TP biểu thị các giá trị dự đoán dương và đúng và FN biểu thị các giá trị dự đoán âm và sai. Độ chính xác được tính toán một cách tương tự. Các chỉ số trung bình vĩ mô tập trung nhiều hơn vào dự đoán nhãn hiếm.

	Range	CAML	CNN	Bi-GRU
d_c	$50 - 500$	50	500	–
k	$2 - 10$	10	4	–
q	$0.2 - 0.8$	0.2	0.2	–
ρ	$0, 0.001, 0.01, 0.1$	0	0	0
η	$0.0001, 0.0003,,$ $0.001, 0.003$	0.0001	0.003	0.003
s	$1 - 4$	–	–	1
v	$32 - 512$	–	–	512

Bảng 1: So sánh tham số của các mô hình

3.4 Kết quả đánh giá

Đánh giá chính của mô hình liên quan đến việc dự đoán bộ mã ICD-9 đầy đủ dựa trên văn bản báo cáo bệnh án trong MIMIC-III. Các kết quả này được thể hiện trong Bảng 4

Mô hình CAML cho kết quả mạnh nhất trên tất cả các thước đo. Mô hình tập trung mang lại những cải tiến đáng kể so với mạng nơ-ron tích tụ "vanila" (CNN). Kiến trúc Bi-GRU lặp lại có thể so sánh với CNN vanila và đường cơ sở hồi quy logistic về cơ bản kém hơn tất cả các kiến trúc thần kinh. Mô hình CNN hoạt động tốt nhất có 9,86M thông số có thể điều chỉnh, so với 6,14M thông số có thể điều chỉnh cho CAML. Điều này là do tìm kiếm siêu tham số thích số lượng bộ lọc lớn hơn cho CNN. Cuối cùng, chúng tôi nhận thấy rằng DR-CAML hoạt động kém hơn trên hầu hết các chỉ số so với CAML, với hệ số điều chỉnh điều chỉnh là $lamda = 0,01$.

Trong những nghiên cứu trước đó, chỉ có mô hình của Scheurwegs **andothers** 2017 đánh giá trên tập dữ liệu MIMIC-III hoàn chỉnh. Kết quả được báo cáo của họ phân biệt giữa mã chẩn đoán và mã thủ tục. Kết quả là mô hình CAML cho kết quả tốt hơn trên cả hai bộ. Ngoài ra, phương pháp đề xuất của tác giả không sử dụng bất kỳ thông tin bên ngoài hoặc dữ liệu có cấu trúc nào, trong khi Scheurwegs **andothers** 2017 sử dụng dữ liệu có cấu trúc và các bản thể học y tế khác nhau trong biểu diễn văn bản của chúng.

Phương pháp đánh giá thứ 2 Ngoài chạy mô hình trên tập dữ liệu chính là MIMIC-III, tác giả có thực hiện trên cả tập dữ liệu 50 mã bệnh thường gặp trong MIMIC-III và trên tập dữ liệu MIMIC-II.

Tác giả thống kê kết quả DR-CAML trên cài đặt 50 nhãn của MIMIC-III với $lamda = 10$ và trên MIMIC-II với $lamda = 0.1$, được xác định bằng tìm kiếm lưới trên bộ xác thực. Các thông số quan trọng khác được giữ nguyên như cài đặt cho đánh giá MIMIC-III chính. Trong cài đặt 50 nhãn của MIMIC-III, chúng tôi thấy sự cải thiện mạnh mẽ so với

	AUC		F1		
Model	Macro	Micro	Macro	Micro	P@ 5
C-MemNN (Prakash et al., 2017)	0.833	–	–	–	0.42
Shi et al. (2017)	–	0.900	–	0.532	–
Logistic Regression	0.829	0.864	0.477	0.533	0.546
CNN	0.876	0.907	**0.576**∗	0.625	**0.620**
Bi-GRU	0.828	0.868	0.484	0.549	0.591
CAML	0.875	0.909	0.532	0.614	0.609
DR-CAML	**0.884**	**0.916**	**0.576**∗	**0.633**	0.618

Bảng 2: Bảng so sánh kết quả đánh giá trên tập dữ liệu 50 nhãn phổ biến của MIMIC-III

	AUC		F1		
Model	Macro	Micro	Macro	Micro	P@ 8
Flat SVM (Perotte et al., 2013)	–	–	–	0.293	–
HA-GRU (Baumel et al., 2018)	–	–	–	0.366	–
Logistic Regression	0.690	0.934	0.025	0.314	0.425
CNN	0.742	0.941	0.030	0.332	0.388
Bi-GRU	0.780	0.954	0.024	0.359	0.420
CAML	0.820	**0.966**∗	0.048	0.442	**0.523**∗
DR-CAML	**0.826**	**0.966**∗	**0.049**	**0.457**∗	0.515

Bảng 3: Bảng so sánh kết quả đánh giá trên tập dữ liệu MIMIC-II

công việc trước đây trong tất cả các chỉ số được báo cáo, cũng như so với các đường cơ sở, ngoại trừ độ chính xác @ 5, mà đường cơ sở CNN hoạt động tốt nhất. Chúng tôi đưa ra giả thuyết rằng điều này là do giá trị tương đối lớn của $lamda = 10$ đối với CAML dẫn đến một mạng lớn hơn phù hợp hơn với các bộ dữ liệu lớn hơn; việc điều chỉnh các siêu tham số của CAML trên tập dữ liệu này sẽ được kỳ vọng sẽ cải thiện hiệu suất trên tất cả các chỉ số.Baumel **andothers** 2017 cũng báo cáo điểm vi mô F1 là 0,407 bằng cách đào tạo về MIMIC-III và đánh giá trên MIMIC-II. Mô hình của chúng tôi đạt được hiệu suất tốt hơn chỉ bằng cách sử dụng bộ đào tạo MIMIC-II (nhỏ hơn), để lại giao thức đào tạo thay thế này cho công việc trong tương lai.

4 Đánh giá tính diễn giải

Tác giả đã đánh giá các đoạn văn bản diễn giải được trích xuất từ cơ chế tập trung của CAML tạo ra, so với ba phương pháp phỏng đoán thay thế.

Phương pháp đánh giá như sau: Tác giả đã trích xuất ra các đoạn văn bản được tập trung bởi mỗi mô hình để cho một bác sĩ đánh giá đoạn văn bản đó có chứa thông tin quan trọng hay không. Người đánh giá sẽ chọn tất cả các đoạn văn bản mà anh ta cảm thấy đã giải thích đầy đủ về sự hiện diện của một mã bệnh nhất định, cung cấp mã và mô tả của nó, với tùy chọn để phân biệt các đoạn trích là "có nhiều thông tin"nếu chúng được tìm thấy đặc biệt nhiều thông tin hơn những đoạn mã khác, "có thông tin"hay đoạn văn bản không liên quan đến kết quả đưa ra.

4.1 Trích xuất các thông tin từ văn bản

- **CAML** Cơ chế tập trung của mô hình cho phép trích xuất k-grams từ những đoạn ghi chứ có nhiều thông tin ảnh hưởng đến dự đoán kết quả của mỗi nhãn, bằng cách lấy argmax (đối số của giá trị cực đại) của đầu ra $\boldsymbol{\alpha}_\ell$.

- **Max-pooling CNN** Tác giả đã chọn ra các đoạn k-grams mà có giá trị lớn nhất trong lớp max-pooling. Và kết quả đầu ra được định nghĩa như sau

$$\boldsymbol{a}_i = \underset{j \in \{1,\ldots,m-k+1\}}{\arg\max} \left(\boldsymbol{H}_{ij}\right),$$

Và chúng ta có thể tính độ quan trọng của vị trí i cho nhãn ℓ

$$\alpha_{i\ell} = \sum_{j:a_j=i}^{d_c} \beta_{\ell,j}$$

Cuối cùng lựa chọn ra k-gram của nhãn kết quả như sau $\arg\max_i \alpha_{i\ell}$.

Method	Informative	Highly informative
CAML	46	22
Code Descriptions	48	20
Logistic Regression	41	18
CNN	36	13

Bảng 4: Bảng kết quả đánh giá tính diễn giải

- **Logistic regression** Tính thông tin của mỗi k-gram thể hiện trong nhãn ℓ được đánh giá bởi tổng của hệ số của đồ thị trong số ℓ, dựa trên số lượng từ trong k-gram. Những k-gram có điểm số cao nhất sẽ được chọn làm văn bản diễn giải.

- **Code descriptions** Tác giả tính toán sự giống nhau giữa mô tả của các mã bệnh đối với kết quả bằng phương pháp idf có trọng số với hàm cosine tương tự. Và lấy ra argmax của k-gram trong đoạn văn bản. Để đảm bảo tính công bằng tác giả đã loại đi các nhãn có điểm không lớn hơn 0 và gây ra việc không thể lựa chọn văn bản cho phương pháp này.

4.2 Kết quả đánh giá

Kết quả đánh giá quá trình diễn giải được trình bày ở bảng dưới đây. Có thể thấy mô hình đề xuất của tác giả đã chọn số lượng lớn nhất các đoạn văn bản diễn giải "có tính thông tin cao" và chọn các giải thích "có thông tin" hơn cả mô hình cơ sở CNN và mô hình hồi quy logistic. Mặc dù Consine Sim cũng hoạt động tốt, các ví dụ trong Bảng 1 chứng minh các điểm mạnh của CAML trong việc trích xuất các đoạn văn bản phù hợp với các giải thích trực quan hơn về sự hiện diện của mã bệnh. Như đã lưu chú ở trên, có một số trường hợp mà chúng tôi loại trừ, trong đó phương pháp tương tự cosin không thể đưa ra bất kỳ lời giải thích nào, bởi vì không có k-gram nào trong ghi chú có sự tương tự khác không đối với mô tả nhãn nhất định. Điều này xảy ra cho khoảng 12% của tất cả các cặp nhãn ghi chú trong bộ thử nghiệm.

Tài liệu tham khảo

Baumel, Tal **andothers** (2017). "Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment". **in***CoRR*: abs/1709.09587. arXiv: 1709.09587. URL: http://arxiv.org/abs/1709.09587.

Scheurwegs, Elyne **andothers** (2017). "Selecting relevant features from the electronic health record for clinical code prediction". **in***Journal of Biomedical Informatics*: 74, **pages** 92–103. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2017.09.004. URL: https://www.sciencedirect.com/science/article/pii/S1532046417302010.

Shi, Haoran **andothers** (2017). "Towards Automated ICD Coding Using Deep Learning". **in***CoRR*: abs/1711.04075. arXiv: 1711.04075. URL: http://arxiv.org/abs/1711.04075.