ĐỀ THI OLPMTTN 2025 BẢNG AI

VÒNG CHUNG KẾT

Vòng đánh giá mở (Public test) : *từ 7h ngày 19-03-2025 đến 7h ngày 20-03-2025*

Vòng đánh giá đóng (Private test) : *từ 8h ngày 20-03-2025 đến 16h ngày 20-03-2025*

- Trang web nộp bài: TẠI ĐÂY
- Mỗi đội chỉ được tạo duy nhất 01 tài khoản trên trang web cuộc thi để nộp bài, tài khoản phải khai báo đầy đủ thông tin như đã đăng ký với BTC.

TỔNG QUAN ĐỀ THI

LĨNH VỰC HỌC MÁY

TÁC VỤ 1: CHẨN ĐOÁN BỆNH DỰA TRÊN HÌNH DẠNG TẾ BÀO TRONG XÉT NGHIỆM

Trong lĩnh vực y tế, các phương pháp chẩn đoán dựa trên hình ảnh vi sinh và tế bào học đóng vai trò quan trọng trong việc phát hiện sớm bệnh tật. Một số bệnh như ung thư, nhiễm trùng, và các rối loạn về máu có thể được xác định thông qua việc phân tích hình dạng và kích thước của tế bào trong các mẫu xét nghiệm.
Hiện nay, quá trình phân tích hình ảnh tế bào vẫn chủ yếu dựa vào các chuyên gia y tế, tốn nhiều thời gian và công sức. Vì vậy, nhiệm vụ của bạn là xây dựng một mô hình trí tuệ nhân tạo (AI) có thể tự động phân loại tế bào thành các nhóm bệnh lý khác nhau dựa trên đặc trưng hình học của chúng.

1.1 Nhiệm vụ

Xây dựng một mô hình trí tuệ nhân tạo (AI) có thể tự động phân loại tế bào thành các nhóm bệnh lý khác nhau dựa trên đặc trưng hình học của chúng.

1.2 Dữ liệu

Dữ liệu được chia thành 4 tập như sau:

- **Public Training set**: Gồm `X_public_train.csv` và `y_public_train.csv`, chứa các đặc trưng hình học của tế bào.
- **Public Test set**: Gồm `X_public_test.csv`, được sử dụng để đánh giá mô hình.
- **Private Training set**: Gồm `X_private_train.csv` và `y_private_train.csv`, chứa các đặc trưng hình học của tế bào.
- **Private Test set**: Gồm `X_private_test.csv`, được Ban tổ chức sử dụng để đánh giá mô hình cuối cùng.

Tập dữ liệu chứa thông tin về hình dạng của các tế bào được trích xuất từ ảnh kính hiển vi. Mỗi hàng đại diện cho một tế bào và được mô tả bởi 22 đặc trưng hình học (cột `f1 - f22`). Cột `label` chứa nhãn bệnh.
Tập dữ liệu kiểm tra sẽ có cùng định dạng với tập huấn luyện nhưng không chứa nhãn (giống X_train).

1.2.1 Cấu trúc file nộp kết quả

File CSV nộp có một cột duy nhất là `label`, chứa nhãn bệnh tương ứng trong tập test (Ví dụ: file `y_public_train.csv` tương ứng với tập `X_public_train.csv`).

Dữ liệu có thể tải về tại: Tải dữ liệu

1.3 Hướng dẫn nộp kết quả

- Không được chỉnh sửa file kết quả do mô hình của đội sinh ra bằng cách gán nhãn thủ công.

- Số lần tối đa được nộp mỗi ngày được quy định trên trang web cuộc thi.
- Các đội lọt vào vòng chung kết bắt buộc phải nộp báo cáo kỹ thuật và source code.
- Mã nguồn phải có đầy đủ các bước xử lý dữ liệu, huấn luyện mô hình, đảm bảo có thể tái hiện kết quả.
- File mã nguồn phải ở định dạng `.ipynb` và có thể chạy trên Colab hoặc Kaggle.

1.4 Tiêu chí đánh giá

Mô hình sẽ được đánh giá dựa trên độ đo **MACRO-F1**. Độ đo này được tính độ đo F1 (F1-Score) riêng cho từng class và sau đó lấy trung bình trên tất cả classes. F1-Score phù hợp với bài toán này vì nó cân bằng giữa precision và recall, đảm bảo đánh giá tin cậy cho các tập dữ liệu không cân bằng.
Công thức để tính F1-Score cho từng class như sau:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

Trong đó:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

LĨNH VỰC THỊ GIÁC MÁY TÍNH

TÁC VỤ 2: PHÂN LOẠI VIÊN THUỐC TRONG HÌNH ẢNH CHẤT LƯỢNG THẤP

Thuốc đóng vai trò quan trọng trong việc hỗ trợ điều trị và cải thiện sức khỏe cho bệnh nhân. Tuy nhiên, việc sử dụng nhầm thuốc có thể dẫn đến những hậu quả nghiêm trọng như giảm hiệu quả điều trị, gây tác dụng phụ nguy hiểm, thậm chí đe dọa tính mạng. Theo Tổ chức Y tế Thế giới (WHO), có đến một phần ba số ca tử vong liên quan đến việc dùng sai thuốc, chứ không phải do bệnh lý gây ra. Trong bối cảnh nhu cầu sử dụng thuốc ngày càng gia tăng và chủng loại thuốc ngày càng đa dạng, việc phát triển các ứng dụng hỗ trợ nhận diện và tra cứu thông tin viên thuốc trở nên vô cùng cần thiết để đảm bảo an toàn cho người dùng.

2.1 Nhiệm vụ

Nhiệm vụ của tác vụ này là xác định tên của các viên thuốc trong một bức ảnh. Cụ thể, đầu vào sẽ là một bức ảnh chụp một số viên thuốc, đầu ra của mô hình sẽ là tên của loại viên thuốc trong ảnh chụp. Tuy nhiên, bộ dữ liệu thu thập được gặp phải nhiều vấn đề về chất lượng hình ảnh chụp. Cụ thể, hình ảnh viên thuốc có thể bị che khuất một phần, mờ, nhiễu hoặc chói sáng, dẫn đến thách thức không nhỏ cho bài toán phân loại viên thuốc trong ảnh. Vì vậy, bạn cần phát triển một mô hình AI mạnh mẽ có khả năng dự đoán tổng quát và chính xác loại viên thuốc trong nhiều trường hợp đầu vào là ảnh chất lượng thấp và cả khi ảnh rõ nét.

2.2 Dữ liệu

Dữ liệu được chia thành 4 tập như sau:

- **Public Training set**: Tập dữ liệu huấn luyện (`train.zip`) chứa các ảnh định dạng JPG mô tả hình ảnh viên thuốc. Bên cạnh đó, các đội được cung cấp file `train_label.csv` mô tả thông tin về nhãn của từng ảnh trong tập huấn luyện (file này có cấu trúc tương tự file kết quả các đội cần nộp).
- **Public Test set**: Tập dữ liệu kiểm thử (`public_test.zip`) bao gồm các ảnh định dạng JPG chưa được gán nhãn, được sử dụng để các đội đánh giá hiệu quả của mô hình mà họ đã huấn luyện.
- **Private Training set**: Tập dữ liệu huấn luyện (`private_train.zip`) chứa các ảnh định dạng JPG mô tả hình ảnh viên thuốc. Bên cạnh đó, các đội được cung cấp file `private_train_label.csv` mô tả thông tin về nhãn của từng ảnh trong tập huấn luyện (file này có cấu trúc tương tự file kết quả các đội cần nộp).
- **Private Test set**: Tập dữ liệu kiểm thử (`private_test.zip`) bao gồm các ảnh định dạng JPG chưa được gán nhãn, được sử dụng để để các đội đánh giá hiệu quả của mô hình mà họ đã huấn luyện.

Hai tập dữ liệu public (Public Training, Public Test) và khung mã nguồn mẫu cung cấp mô hình (Python, Keras) được chia sẻ qua đường link NÀY. Tại vòng Public Test, số lượng loại viên thuốc khác nhau là **10**. Các đội cần sử dụng mô hình được cung cấp sẵn (trong cell có chú thích về mô hình không được sửa đổi) và có thể tham khảo code huấn luyện và sinh output của BTC.

2.3 Hướng dẫn nộp kết quả

- Không được chỉnh sửa file kết quả do mô hình của đội sinh ra bằng cách gán nhãn thủ công các dữ liệu trong tập test (bao gồm public test lẫn private test) để nộp cho BTC.
- Số lần tối đa được nộp trong một ngày của Giai đoạn sơ khảo (public test) và Giai đoạn chung kết (private test) được quy định trên trang web cuộc thi.
- Các đội lọt vào vòng chung kết bắt buộc phải nộp báo cáo kỹ thuật và source code để BTC đánh giá tính đúng đắn của giải pháp và công bố kết quả cuối cùng.
- Các đội nộp một file mã nguồn colab hoặc kaggle (đuôi `.ipynb`), trong đó có đầy đủ các bước cài đặt môi trường, xử lý dữ liệu, huấn luyện mô hình, ... theo đúng thứ tự để BTC có thể reproduce được các mô hình.
- Mã nguồn của các đội cần tuân thủ chặt chẽ định dạng dữ liệu đầu vào và kết quả trả ra để BTC sẽ chạy mã nguồn của các đội, đánh giá chất lượng (dựa trên các tiêu chí tối ưu đề ra) của kết quả đầu ra và xếp hạng.

2.4 Quy định

Đội thí sinh không được phép thay đổi các cấu hình tham số sẵn có của mô hình. Đối với tác vụ này, mô hình cần được huấn luyện từ đầu (tức không sử dụng pre-trained weights). Cần đảm bảo rằng BTC có thể tải checkpoint weights do đội thí sinh cung cấp và chạy ra kết quả ứng với điểm số trên leaderboard.

2.5 Tiêu chí đánh giá

Mô hình sẽ được đánh giá dựa trên độ đo **MACRO-F1**.

LĨNH VỰC XỬ LÝ NGÔN NGỮ TỰ NHIÊN

TÁC VỤ 3: KHÔI PHỤC VĂN BẢN ĐƯỢC MÃ HÓA

Công ty MTTN nhận được một yêu cầu từ một thư viện nổi tiếng của quốc gia. Ở thư viện, có một số sách hiếm đã bị xuống cấp nên một số từ không còn nhìn được rõ. Do đó, nhiệm vụ thủ thư đặt ra cho công ty là dựa vào một số văn bản tương tự đã số hóa hãy tìm cách khắc phục lại những từ bị thiếu kia. Tuy vậy, do lí do về bảo mật, các đội thi sẽ không được cung cấp các văn bản này với nội dung nguyên gốc của chúng. Thay vào đó, các từ trong tập văn bản sẽ được số hóa hoàn toàn, và các đội thi có nhiệm vụ phải dự đoán các từ bị thiếu dưới dạng các số đại diện cho từ đó (indices).

3.1 Nhiệm vụ

Bạn được cung cấp một tập dữ liệu bao gồm các văn bản hoàn thiện của thư viện, với đầy đủ nội dung được mã hóa. Từ tập dữ liệu này, nhiệm vụ của bạn là xây dựng chương trình có khả năng đoán những từ còn thiếu trong các văn bản với từ bị thiếu một cách tốt nhất, với các ràng buộc sau:

- Mô hình phải đào tạo trên môi trường colab và kaggle để có thể chạy lại được.
- Không được sử dụng bất kỳ dữ liệu nào khác ngoài dữ liệu trong tập Training set cũng như bất kỳ mô hình đã huấn luyện sẵn (pretrained models).

3.2 Dữ liệu

Dữ liệu gồm 4 tập:

- **Training / Validation set**: Chứa tập ngữ liệu văn bản (`train.json / validation.json`) dùng để huấn luyện.
- **Public Test set**: Tập dữ liệu kiểm thử công khai (`public_test.pkl`), được sử dụng để các đội đánh giá hiệu quả của mô hình mà họ đã huấn luyện.
- **Private Training set**: Tập dữ liệu huấn luyện cho vòng Private Test.
- **Private Test set**: Tập dữ liệu kiểm thử ẩn, được Ban tổ chức sử dụng để đánh giá độ chính xác của các mô hình từ các đội thi.

Hai tập dữ liệu training và public test được chia sẻ qua đường link NÀY.

Ngoài ra, các đội thi cũng được cung cấp 2 notebook mã nguồn mẫu `torch-notebook-final.ipynb` và `keras-notebook-final.ipynb`

3.3 Hướng dẫn nộp kết quả

- Không được chỉnh sửa file kết quả do mô hình của đội sinh ra bằng cách gán nhãn thủ công các dữ liệu trong tập test (bao gồm public test lẫn private test) để nộp cho BTC.
- Giai đoạn sơ khảo (public test): mỗi đội được phép nộp tối đa 15 lần mỗi ngày. Trong giai đoạn chung kết (private test) được phép nộp tối đa 15 lần mỗi ngày. Các đội lọt vào vòng chung kết bắt buộc phải nộp báo cáo kỹ thuật và source code để BTC đánh giá tính đúng đắn của giải pháp và công bố kết quả cuối cùng.

- Các đội nộp file mã nguồn notebook (đuôi `.ipynb`), trong đó có đầy đủ các bước cài đặt môi trường, xử lý dữ liệu, huấn luyện mô hình, ... theo đúng thứ tự để BTC có thể reproduce lại được các mô hình.
- Mã nguồn của các đội cần tuân thủ chặt chẽ định dạng dữ liệu đầu vào và kết quả trả ra để BTC sẽ chạy mã nguồn của các đội, đánh giá chất lượng (dựa trên các tiêu chí tối ưu đề ra) của kết quả đầu ra và xếp hạng.

3.4 Tiêu chí đánh giá

Mô hình sẽ được đánh giá dựa trên độ đo **Brier score**. Điểm Brier là một quy tắc chấm điểm hoàn toàn chính xác dùng để đo độ chính xác của các dự đoán xác suất. Đối với các dự đoán một chiều, nó tương đương với lỗi bình phương trung bình khi áp dụng cho các xác suất dự đoán. Công thức của chúng được thể hiện như sau:

$$BS = \frac{1}{N}\sum_{i=1}^{N}\left(f_t - o_t\right)^2. \tag{4}$$

trong đó f_t là xác suất đã được dự báo được xuất ra bởi đầu ra của mô hình, o_t kết quả thực tế của sự kiện tại mẫu t và N là số trường hợp dự báo.