ĐỀ THI OLPMTTN 2025 BẢNG AI

VÒNG SƠ LOẠI

Vòng đánh giá mở (Public test) : *từ 16-02-2025 đến 23-02-2025*

Vòng đánh giá đóng (Private test) : *từ 8h ngày 23-02-2025 đến 16h ngày 23-02-2025*

- 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: Dự đoán Mức Độ Hao Mòn của Máy Móc Công Nghiệp

Một tập đoàn sản xuất thiết bị công nghiệp hàng đầu đang tìm cách nâng cao hiệu suất vận hành bằng cách phân tích mức độ hao mòn của máy móc trong dây chuyền sản xuất. Họ đã thu thập một lượng lớn dữ liệu kỹ thuật về các thiết bị, nhưng một số trường dữ liệu bị thiếu giá trị do hạn chế trong quá trình thu thập.

Để tối ưu hóa lịch trình bảo trì và giảm thiểu rủi ro hỏng hóc đột xuất, công ty cần một mô hình máy học có khả năng ước lượng chính xác mức độ hao mòn của từng thiết bị. Mô hình này không chỉ giúp cải thiện độ tin cậy của hệ thống mà còn tối ưu chi phí bảo trì và kéo dài tuổi thọ của máy móc.

1.1 Nhiệm vụ

Xây dựng một mô hình dự đoán mức độ hao mòn của thiết bị dựa trên các đặc điểm kỹ thuật và điều kiện vận hành. Mô hình cần xử lý dữ liệu bị nhiễu hoặc thiếu và có khả năng cung cấp dự đoán chính xác để hỗ trợ tối ưu hóa bảo trì.

1.2 Dữ liệu

Dữ liệu được chia thành 3 tập như sau:

- **Training set:** Tập dữ liệu huấn luyện (`X_train.csv` và `y_train.csv`), chứa thông tin về các thiết bị công nghiệp, bao gồm thông số kỹ thuật, lịch sử bảo trì, và dữ liệu từ cảm biến và mức độ hao mòn của thiết bị.
- **Public testset:** Tập dữ liệu kiểm thử công khai (`X_public_test.csv`), được sử dụng để các đội đánh giá hiệu quả của mô hình mà họ đã huấn luyện.
- **Private testset:** 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.

Tập dữ liệu này chứa thông tin về thiết bị công nghiệp và mức độ hao mòn của chúng. Dữ liệu đã được chuẩn hóa và gồm hai nhóm chính:

- **Các trường numerical:** Bao gồm các chỉ số kỹ thuật và thông tin vận hành, được chuẩn hóa và đánh số lần lượt như `num1, num2, ..., num12`. Một số trường có giá trị thiếu do lỗi cảm biến hoặc dữ liệu không đầy đủ.
- **Các trường categorical:** Biểu diễn thông tin phân loại về thiết bị, môi trường hoạt động hoặc lịch sử bảo trì, được mã hóa dưới dạng số nguyên như `cat1, cat2, ..., cat8`.

Tập dữ liệu huấn luyện X_train có tổng cộng 20 cột với 17.479 mẫu. Cột về độ hao mòn tương ứng đã được tách ra thành y_train.

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

Nộp file CSV có một cột duy nhất là wear_rate là độ hao mòn tương ứng với tập test (giống y_train)

Dữ liệu được download:

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 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: TẠI ĐÂY.
- 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.

1.4 Tiêu chí đánh giá

Đánh giá mô hình dựa trên độ chính xác dự đoán bằng RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

Công thức này đo lường sai số trung bình bình phương giữa giá trị thực tế y_i và giá trị dự đoán $\hat{y}_i$ trên n mẫu thuộc tập test.

LĨNH VỰC THỊ GIÁC MÁY TÍNH

TÁC VỤ 2: PHÂN LOẠI HÌNH ẢNH GIẢI PHẪU TẾ BÀO

Phân tích hình ảnh giải phẫu tế bào trên kính hiển vi đóng vai trò quan trọng trong việc phát hiện sớm và chẩn đoán nhiều loại bệnh lý. Tuy nhiên, đánh giá thủ công tốn nhiều thời gian, phụ thuộc vào chuyên gia và thường có nguy cơ sai sót. Vì vậy, bằng cách ứng dụng công nghệ học máy và thị giác máy tính, bài toán này hướng đến việc nâng cao độ chính xác và hiệu suất trong phân tích hình ảnh giải phẫu tế bào, đặc biệt hữu ích cho các khu vực bệnh viện có hạn chế về nhân lực và trang thiết bị y tế.

2.1 Nhiệm vụ

Bạn được cung cấp một tập dữ liệu chuyên sâu, được gán nhãn bởi các chuyên gia, bao gồm hàng nghìn hình ảnh hiển vi số hóa. Những hình ảnh này được phân loại thành ba nhóm (xem Hình 1), trong đó gồm:

- **Healthy**: Mẫu có đặc điểm tiêu chuẩn, không có dấu hiệu bất thường.
- **Unhealthy**: Mẫu có sự biến đổi về mặt cấu trúc, có thể liên quan đến nguy cơ bệnh lý.
- **Rubbish**: Hình ảnh bị nhiễu, mờ hoặc không đạt chất lượng để đánh giá chính xác.

Một trong những thách thức lớn của bài toán này là sự mất cân bằng dữ liệu, khi số lượng mẫu bất thường thấp hơn đáng kể so với mẫu bình thường và mẫu không thể sử dụng. Điều này phản ánh thực tế lâm sàng, đòi hỏi các giải pháp kỹ thuật vững chắc để đảm bảo mô hình phân loại chính xác trên toàn bộ tập dữ liệu.

Bạn sẽ cần phát triển một mô hình có khả năng phân biệt chính xác giữa ba nhóm mẫu hình ảnh, vượt qua các thách thức về sự mất cân bằng, sự đa dạng trong hình thái mẫu quan sát và các yếu tố gây nhiễu trong quá trình thu nhận dữ liệu.

Hình 1: Mô tả về các mẫu hình ảnh hiển vi trong tập dữ liệu

2.2 Cấu trúc dữ liệu

Dữ liệu được chia thành 3 tập như sau:

- **Training set:** Tập dữ liệu huấn luyện (`train.zip`) chứa 3 folder lần lượt là `healthy`, `unhealthy` và `rubbish`. Mỗi folder chứa các ảnh định dạng PNG mô tả mẫu dữ liệu tương ứng như đã mô tả.

- **Public testset:** Tập dữ liệu kiểm thử công khai (`public_test.zip`) bao gồm các ảnh định dạng PNG chưa được gán nhãn, được sử dụng để các đội đánh giá hiệu quả của mô hình mà họ đã huấn luyện.
- **Private testset:** Tập dữ liệu kiểm thử ẩn (`private_test.zip`), được Ban tổ chức sử dụng để đánh giá độ chính xác của các mô hình từ các đội thi.

2.2.1 Cấu trúc dữ liệu file nộp kết quả

File kết quả là một file CSV có định dạng như sau:

image_name	label
image_001.png	healthy
image_002.png	unhealthy
image_003.png	rubbish

Trong đó, **image_name** là cột chứa tên của hình ảnh và **label** là cột chứa nhãn tương ứng do mô hình dự đoán (healthy, unhealthy, rubbish).

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: TẠI ĐÂY.
- 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 Tiêu chí đánh giá

Kết quả dự đoán của các đội được đánh giá theo độ đo **F1-Score**. Độ đo này được tính 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:

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

Trong đó:

Precision:

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

Recall:

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

LĨNH VỰC XỬ LÝ NGÔN NGỮ TỰ NHIÊN

TÁC VỤ 3: PHÂN LOẠI CẢM XÚC TRONG VĂN BẢN

Công ty MTTN được một khách hàng yêu cầu phân tích cảm xúc từ văn bản tiếng Anh. Việc phân tích cảm xúc này giúp cho công ty đánh giá được chất lượng chăm sóc dịch vụ trong các cuộc hội thoại giữa nhân viên và khách hàng.

Do yêu cầu của việc đưa ứng dụng vào điện thoại nên kích thước mô hình cần có kích thước nhỏ. Thêm nữa để đảm bảo vấn đề về bản quyền, công ty đưa ra yêu cầu cho bạn là phải huấn luyện mô hình từ đầu và không được sử dụng bất kỳ mô hình có tham số bên ngoài vào bất kỳ bước nào kể cả quá trình tiền xử lý, huấn luyện hay suy diễn mô hình.

3.1 Nhiệm vụ

Nhiệm vụ của bạn là xây dựng trình phân loại cảm xúc trong văn bản tốt nhất, với các ràng buộc sau:

- Mô hình phải đào tạo trong vòng chưa đầy 1 giờ bằng cách sử dụng GPU L4 vì tài nguyên tính toán của công ty còn hạn chế.
- Kích thước của mô hình (số khích thước bộ nhớ của tham số) không được vượt quá 10 MB.
- 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 Cấu trúc dữ liệu

Dữ liệu được chia thành 3 tập như sau:

- **Training set:** Tập dữ liệu huấn luyện (`train_X.pkl` và `train_y.pkl`), chứa danh sách các văn bản và nhãn tương ưng của chúng. Các nhãn có ý nghĩa: 0 - buồn bã, 1 - vui vẻ, 2 - yêu thích, 3 - giận dữ, 4 - sợ hãi, 5 - bất ngờ.
- **Public testset:** Tập dữ liệu kiểm thử công khai (`valid_X.pkl`), được sử dụng để các đội đánh giá hiệu quả của mô hình mà họ đã huấn luyện.
- **Private testset:** 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.

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.
- 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: TẠI ĐÂY.
- 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.

3.4 Tiêu chí đánh giá

Nhãn dự đoán của các đội sẽ được so sánh với nhãn thực tế và tính theo độ đo Marco-F1. Cụ thể công thức đánh giá được tính như sau:

$$Marco - F1 = \frac{1}{6}\sum_{k=1}^{6} F1_k \tag{2}$$

Trong đó: $F1_k$ là điểm $F1$ của nhãn thứ k.