

BÁO CÁO PHÂN TÍCH KẾT QUẢ HUẤN LUYỆN

Dự đoán hỏng hóc máy bơm công nghiệp – 62FIT4ATI

1. Tổng quan dự án

1.1. Bối cảnh và mục tiêu

Máy bơm công nghiệp là thiết bị quan trọng trong nhiều hệ thống sản xuất; sự cố đột ngột có thể gây dừng dây chuyền, tăng chi phí sửa chữa và tiềm ẩn rủi ro an toàn.

Mục tiêu của dự án là xây dựng mô hình mạng nơ-ron hồi quy LSTM để dự đoán trạng thái hoạt động của máy bơm (NORMAL, BROKEN, RECOVERING) dựa trên dữ liệu cảm biến chuỗi thời gian, phục vụ bài toán bảo trì dự đoán. Đồng thời cũng là cơ hội học tập, tiếp cận LSTMs dưới dạng model dùng trong thực tế so với lý thuyết

1.2. Thông tin dữ liệu

Dữ liệu được cung cấp dưới dạng file `sensor.csv`, được đọc trực tiếp trong notebook.

- Số dòng dữ liệu: khoảng 220,320 bản ghi.
 - Số lượng cảm biến: 52 tín hiệu liên tục, kèm theo các cột thời gian và nhãn trạng thái máy.
 - Sau bước cắt chuỗi và chuẩn hóa, thu được 22,027 chuỗi thời gian độ dài 50 timestep, mỗi timestep có 51 đặc trưng đầu vào.
 - Bài toán là phân loại 3 lớp:
 - o NORMAL: bơm hoạt động bình thường
 - o BROKEN: bơm hỏng
 - o RECOVERING: bơm trong giai đoạn phục hồi sau sự cố hoặc bảo trì
 - Phân bố nhãn mất cân bằng mạnh, các lớp hỏng và phục hồi chiếm tỉ lệ nhỏ so với lớp bình thường.
 - Dùng 3 lớp để phân biệt, đồng thời cũng để biến đổi sao cho độ chuẩn xác cao hơn
-

2. Kiến trúc và cấu hình mô hình

2.1. Kiến trúc mô hình

Mô hình được xây dựng trong notebook `62FIT4ATI_Group_12_Topic_2.ipynb` dưới dạng mạng LSTM nhiều tầng.

- Loại: mạng nơ-ron hồi quy LSTM cho time-series classification.
- Số lớp: 9 lớp (2 lớp LSTM, các lớp Batch Normalization, Dropout và Dense).
- Tổng số tham số trainable: 144,515.

- Đầu vào: tensor kích thước (batch_size, 50, 51).
- Đầu ra: vector xác suất gồm 3 phần tử ứng với NORMAL, BROKEN và RECOVERING.

Hai lớp LSTM xếp chồng giúp mô hình học được cả các quan hệ ngắn hạn và dài hạn trong chuỗi cảm biến, trong khi các lớp Batch Normalization và Dropout được sử dụng để ổn định quá trình huấn luyện và giảm overfitting. Nhờ việc giảm overfitting mà có thể giảm thiểu mức độ gây hại cho máy chủ test hay thời gian test khi bắt đầu chạy

2.2. Cấu hình huấn luyện

Các bước cài đặt và huấn luyện được thực hiện trong Colab với GPU.

- Hàm tối ưu: Adam với learning rate khởi điểm 1.00e-04.
 - Hàm mất mát: sparse_categorical_crossentropy do nhãn được mã hóa dạng số nguyên.
 - Các metric theo dõi trong quá trình huấn luyện: accuracy trên tập huấn luyện và tập validation.
 - Dữ liệu được chia thành train, validation và test, sau đó chuẩn hóa bằng StandardScaler và nhãn được mã hóa bằng LabelEncoder.
-

3. Kết quả hiệu suất mô hình

3.1. Hiệu suất trên tập huấn luyện và validation

Trong quá trình huấn luyện, mô hình nhanh chóng cải thiện độ chính xác sau vài epoch đầu.

Epoch tốt nhất trên validation (Epoch 15):

Metric	Training	Validation
Accuracy	99.91%	99.91%
Loss	0.006002	0.006305

Tại Epoch 30, khi kết thúc huấn luyện:

Metric	Training	Validation
Accuracy	99.99%	99.91%
Loss	0.000724	0.009400
LearningRate	2.50e-05	-

Độ chính xác ban đầu ở Epoch 1 chỉ khoảng 41.30%, sau đó tăng lên 99.99% ở Epoch 30, tương ứng mức cải thiện hơn 58%.

Training loss giảm từ 0.846 xuống 0.000724, trong khi validation loss giảm từ 0.204 xuống 0.009400, cho thấy mô hình hội tụ tốt trên cả train và validation.

3.2. Hiệu suất trên tập kiểm thử

Mô hình sau khi huấn luyện được đánh giá trên tập kiểm thử độc lập.

- Test Accuracy: 99.98%
- Test Precision: 0.9998
- Test Recall: 0.9998
- Macro F1-Score: 0.9991
- Weighted F1-Score: 0.9998

Các chỉ số precision và recall rất cao cho thấy mô hình vừa phát hiện tốt các trường hợp hỏng hóc, vừa hạn chế báo động sai.

Sự chênh lệch nhỏ giữa F1 macro và F1 weighted gợi ý rằng ngay cả các lớp hiếm cũng được mô hình học tương đối cân bằng.

4. Phân tích quá trình huấn luyện

4.1. Lịch trình learning rate

Notebook sử dụng callback ReduceLROnPlateau để tự động giảm learning rate khi validation loss không còn cải thiện.

- LR = 1.00e-04 từ Epoch 1 đến 22: giai đoạn học chính, độ chính xác tăng nhanh.
- LR = 5.00e-05 từ Epoch 23 đến 29: giai đoạn tinh chỉnh, giúp bước cập nhật nhỏ hơn và ổn định hơn.
- LR = 2.50e-05 tại Epoch 30: fine-tuning cuối cùng quanh nghiệm tối ưu.

Chiến lược này giúp kết hợp được tốc độ hội tụ nhanh ở giai đoạn đầu với sự ổn định ở giai đoạn sau.

4.2. Đánh giá hiện tượng overfitting

Độ chênh cuối giữa train và validation accuracy chỉ khoảng 0.08%, và trung bình trong 5 epoch cuối khoảng 0.07%.

Khoảng cách nhỏ giữa hai đường cong accuracy và loss trên train và validation cho thấy mô hình không gặp overfitting nghiêm trọng, các kỹ thuật regularization đang hoạt động hiệu quả. Vì vậy chúng ta cần phải một số kỹ thuật để tối ưu, ngoài việc giảm epoch test cũng như tăng mức độ overfitting rate lên

5. Các kỹ thuật tối ưu hóa đã áp dụng

Để mô hình ổn định hơn trên dữ liệu công nghiệp có độ nhiễu và mất cân bằng lớp, nhiều kỹ thuật tối ưu hóa đã được sử dụng kết hợp.

1. Class Weights
 - o Tự động tính trọng số cho từng lớp dựa trên tần suất xuất hiện, giúp các lớp hiếm như BROKEN và RECOVERING được chú ý hơn trong quá trình học.
2. Learning Rate Scheduling với ReduceLROnPlateau

- Giảm learning rate mỗi khi validation loss dừng cải thiện, với hệ số 0.5 và patience 5 epoch, tránh tình trạng học quá thô ở giai đoạn cuối.
3. Early Stopping
 - Dừng huấn luyện nếu mô hình không cải thiện trong 15 epoch liên tiếp và khôi phục bộ trọng số tốt nhất, tiết kiệm thời gian và tránh overfitting.
 4. Dropout Regularization
 - Áp dụng dropout với các tỉ lệ từ 0.2 đến 0.4 sau các lớp LSTM và Dense, giảm phụ thuộc vào một nhóm neuron cụ thể.
 5. Batch Normalization
 - Chuẩn hóa activation giữa các lớp, giúp gradient ổn định hơn và cho phép dùng learning rate ban đầu tương đối cao.
 6. Gradient Clipping
 - Sử dụng clipnorm = 1.0 để giới hạn độ lớn gradient, đặc biệt hữu ích cho các mô hình LSTM xử lý chuỗi dài nhằm tránh exploding gradients.

Sự kết hợp của các kỹ thuật trên là một trong những lý do khiến mô hình vừa đạt hiệu suất cao, vừa giữ được khả năng tổng quát hóa tốt trên tập test.

6. Đánh giá tổng quan và ứng dụng thực tế

6.1. Điểm mạnh của mô hình

Các kết quả thực nghiệm cho thấy:

- Độ chính xác trên validation và test đều lớn hơn 99%, với loss thấp, chứng tỏ mô hình rất phù hợp với tập dữ liệu hiện tại.
- Khoảng cách nhỏ giữa train và validation cho thấy mô hình không bị overfitting đáng kể.
- Các chỉ số F1 macro và weighted cao chứng minh mô hình xử lý khá tốt vấn đề mất cân bằng lớp.
- Quá trình huấn luyện diễn ra ổn định, không ghi nhận hiện tượng loss dao động mạnh hoặc gradient bất thường.

6.2. Khả năng ứng dụng trong nhà máy

Mô hình có thể được tích hợp vào hệ thống giám sát bơm công nghiệp để:

- Dự đoán sớm trạng thái BROKEN, cho phép lên kế hoạch dừng máy và bảo trì chủ động, giảm thời gian ngừng hoạt động ngoài ý muốn.
- Xây dựng lịch bảo trì dựa trên tình trạng thực của thiết bị, thay vì chỉ dựa trên số giờ chạy hoặc chu kỳ cố định.

- Theo dõi trạng thái RECOVERING sau bảo trì để đánh giá chất lượng sửa chữa và điều chỉnh chế độ vận hành.
 - Lưu trữ lịch sử trạng thái để phân tích xu hướng hỏng hóc, tối ưu chiến lược vận hành toàn hệ thống.
-

6.3 Điều rút ra được từ mô hình

Nhờ luyện tập với LSTMs mà bọn em rút ra được một số vấn đề

- Nên chú trọng vào việc xử lý dữ liệu: Dữ liệu khi được xử lý tốt, không có giá trị vô biến(null) hoặc một số giá trị không mong muốn sẽ giúp cho việc triển khai model ổn định hơn, tránh rủi ro sót các chỉ số trong 1 số trường hợp nhất định
- Đừng chỉ nhìn vào độ chính xác mà còn phải xem trong điều kiện khác nhau thì model có sự biến động nào hay không

7. Hướng phát triển tiếp theo

7.1. Cải thiện kiến trúc mô hình

Một số hướng mở rộng có thể nghiên cứu trong tương lai:

- Sử dụng Bidirectional LSTM để mô hình hóa quan hệ theo cả hai chiều thời gian.
- Bổ sung attention mechanism để mô hình tập trung hơn vào các thời điểm quan trọng trong chuỗi.
- Xây dựng các mô hình ensemble kết hợp LSTM với GRU hoặc CNN 1D để tăng độ ổn định.

7.2. Cải thiện chất lượng dữ liệu

- Thu thập thêm dữ liệu thực tế cho các trạng thái ít xuất hiện, đặc biệt là RECOVERING.
- Áp dụng các kỹ thuật data augmentation cho chuỗi thời gian nhằm giảm mất cân bằng lớp.
- Thực hiện feature engineering với các đặc trưng thống kê như trung bình trượt, phương sai, độ dốc và các đặc trưng miền tần số.

7.3. Hướng triển khai thực tế

- Triển khai online learning hoặc cập nhật mô hình định kỳ khi dữ liệu mới được thu thập từ hệ thống.
 - Xây dựng hệ thống giám sát hiệu năng mô hình theo thời gian để phát hiện sớm khi performance suy giảm.
 - Kết hợp các phương pháp giải thích mô hình như SHAP để giúp kỹ sư hiểu rõ hơn lý do mô hình đưa ra dự đoán hỏng hóc.
-

8. Kết luận

Dự án đã xây dựng thành công mô hình LSTM cho bài toán dự đoán hỏng hóc máy bơm công nghiệp với độ chính xác rất cao trên tập kiểm thử, đồng thời xử lý tốt vấn đề mất cân bằng lớp và tránh được overfitting nghiêm trọng.

Mô hình đạt 99.98% accuracy trên tập test, thể hiện tiềm năng lớn để triển khai trong các hệ thống bảo trì dự đoán thực tế, giúp giảm chi phí, nâng cao an toàn và tối ưu vận hành nhà máy.

Các kinh nghiệm chính rút ra bao gồm tầm quan trọng của tiền xử lý dữ liệu time-series, lựa chọn kiến trúc LSTM phù hợp và kết hợp nhiều kỹ thuật tối ưu hóa trong quá trình huấn luyện.