BÁO CÁO DỰ ÁN PHÂN LOẠI ẢNH BẰNG VISION TRANSFORMER

Thực hiện bởi: [Nguyễn Vân]

July 16, 2025

1 Giới thiệu tổng quan

Sử dụng mô hình Vision Transformer (ViT) để thực hiện phân loại ảnh. Cụ thể là cho bài toán phân loại ảnh chứa công trường và không chứa công trường.

2 Mục tiêu

Mục tiêu chính là:

- Ứng dụng mô hình ViT để phân loại một tập ảnh cụ thể.
- Huấn luyện mô hình trên tập huấn luyện, đánh giá trên tập kiểm tra.
- Hiển thị các chỉ số như độ chính xác, F1-score và trực quan hóa confusion matrix.
- Tạo khả năng dự đoán ảnh từ bên ngoài do người dùng đưa vào.

3 Các bước thực hiện

3.1 Tiền xử lý dữ liệu

- Ảnh được resize về kích thước 224x224.
- Sử dụng `AutoImageProcessor` để chuẩn hóa ảnh đầu vào và tạo `pixel_values`.
- Dataset được ánh xạ lại để chứa trường `pixel_values` và `labels`.

3.2 Cài đặt mô hình

- Sử dụng `ViTForImageClassification` từ thư viện `transformers`.
- Khởi tạo Trainer từ Hugging Face với các tham số phù hợp
- Tùy chỉnh `data_collator` để loại bỏ trường không cần thiết như `id`.

3.3 Huấn luyện và đánh giá mô hình

- Mô hình được huấn luyện trong 3 epoch, ghi lại các chỉ số theo từng bước, với `eval_steps` và `save_steps` = 500.
- Đánh giá dựa trên loss, accuracy và F1-score.
- Hiển thị **confusion matrix** sau khi huấn luyện.

3.4 Dự đoán ảnh người dùng đưa vào

- Cho phép người dùng cung cấp ảnh mới từ máy tính.
- Ảnh được xử lý giống như ảnh huấn luyện.
- In ra dự đoán tên lớp của ảnh đó.

4 Kết quả đạt được

Chỉ số đánh giá mô hình sau huấn luyện:

- **Accuracy**: khoảng 97.2%
- **F1-score**: khoảng 97.1%
- **Eval loss**: khoảng 0.14

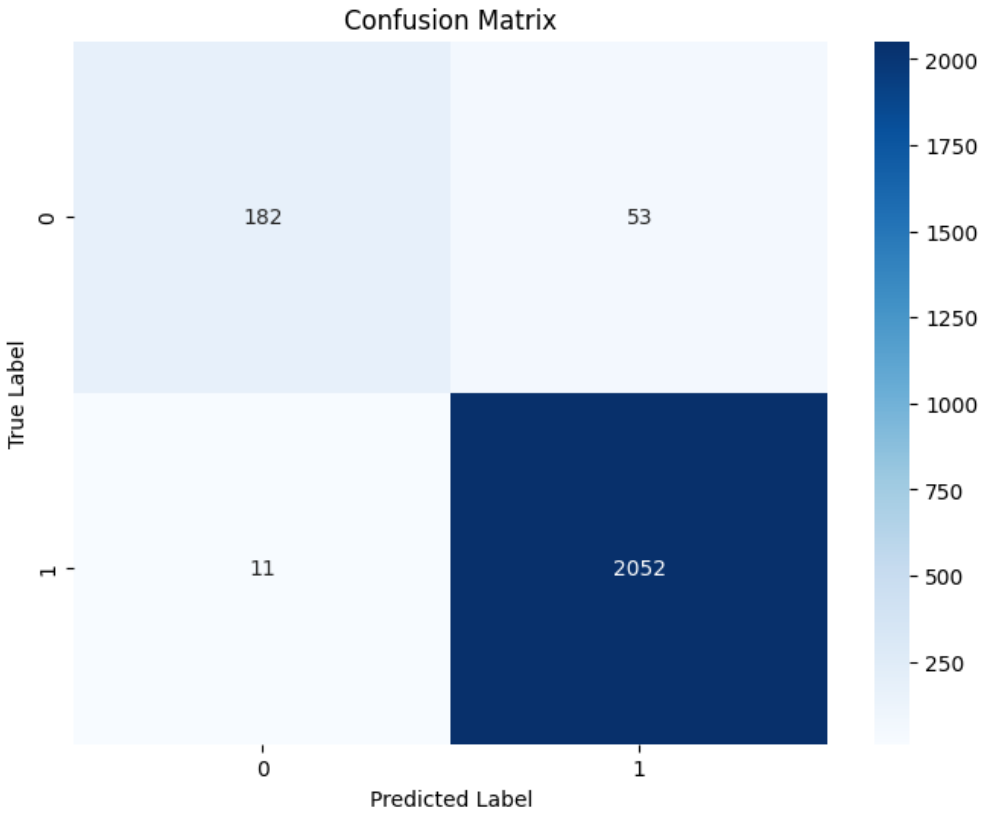
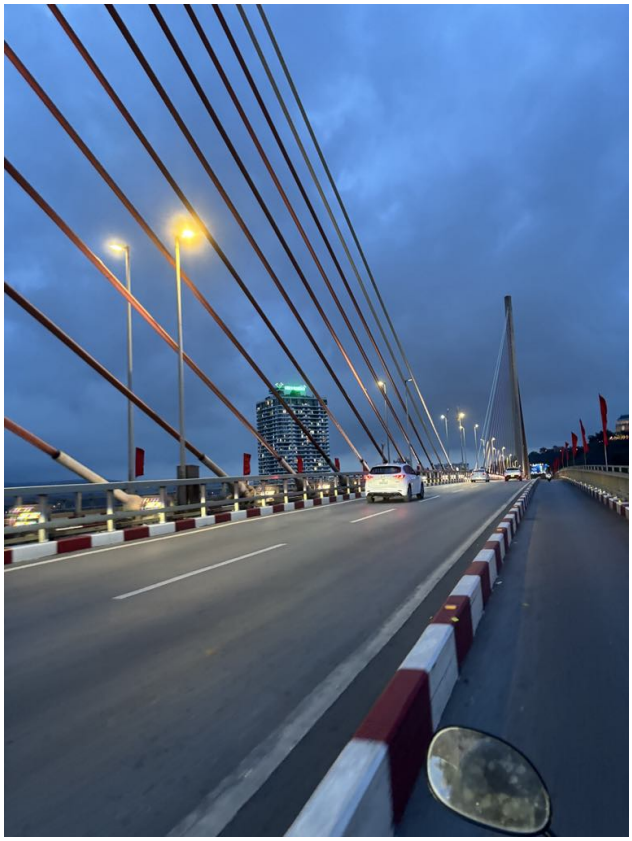


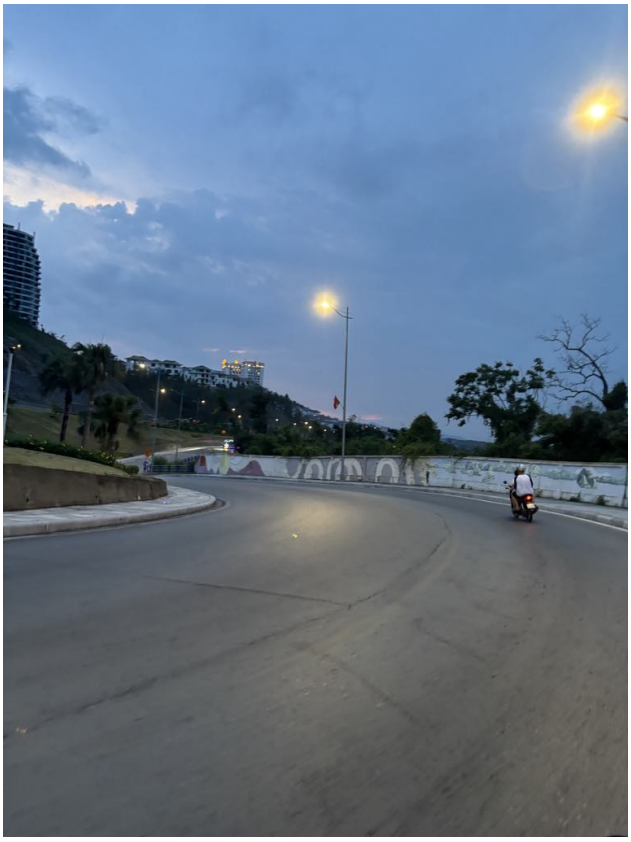
Figure 1: Confusion Matrix trên tập kiểm tra

Thử nghiệm

Dự đoán chính xác với ảnh ban ngày, màu sắc rõ nét:

Dự đoán sai với hai ảnh buổi tối:





5 Những gì đã làm được

- Hoàn thiện pipeline huấn luyện ảnh phân loại bằng ViT.
- Áp dụng thành công mô hình học sâu với kiến trúc hiện đại.
- Trực quan hóa được confusion matrix.
- Viết thêm chức năng dự đoán ảnh mới từ người dùng.
- Mô hình đạt độ chính xác và F1-score cao.

6 Hạn chế và hướng phát triển

Hạn chế

- Chưa vẽ biểu đồ trực quan loss và accuracy theo epoch.
- Chưa áp dụng kỹ thuật giảm overfitting như: early stopping, weight decay, data augmentation.
- Việc dự đoán ảnh mới còn thủ công, chưa có giao diện người dùng.
- Confusion matrix có thể mất thời gian với dataset lớn.

Hướng phát triển

- Tích hợp GUI đơn giản (streamlit, gradio).
- So sánh với mô hình nhẹ hơn (MobileNet, ResNet,...).
- Tăng cường dữ liệu đầu vào để cải thiện độ tổng quát.
- Tự động lưu mô hình tốt nhất trong quá trình huấn luyện.