Report FastAPI with Trism

Report day: 25/02/2025

Name: Nguyễn Thanh Hòa

Team: 2 Team leader: Hòa

Team member: Nguyễn Thanh Hòa, Nguyễn Đình Vũ Hải, Nguyễn Quý Toàn

Team member

✅ **Task title 1: Thay thế tritonclient bằng trism và cải thiện cấu trúc serving model**

▼ 1. Purpose

- Thay thế thư viện `tritonclient` bằng `trism` để đơn giản hóa việc gọi inference từ Triton Inference Server.
- Sử dụng repository `hieupth/tritonserver` để triển khai Triton Inference Server với image nhẹ hơn.
- Lưu trữ model trên Hugging Face và tự động tải xuống mỗi lần chạy Docker image hosting model.
- Cải thiện cấu trúc mô hình trên Hugging Face để phù hợp với Triton Server.

▼ 2. Action

- Cài đặt thư viện `trism` và kiểm thử việc gọi inference từ Triton Inference Server.
- Chỉnh sửa cấu hình Dockerfile để sử dụng image `hieupth/tritonserver`.
- Thiết lập pipeline lưu model trên Hugging Face và tự động tải xuống khi khởi động container.
- Chạy thử nghiệm để kiểm tra tính tương thích.

▼ 3. Result

- Thư viện `trism` đã thay thế `tritonclient` thành công, giúp đơn giản hóa việc gọi inference.
- Image `hieupth/tritonserver` hoạt động tốt, giảm dung lượng container so với image mặc định của Triton.
- Model đã được lưu lên Hugging Face và tải xuống tự động khi chạy Docker container.
- Lỗi:
- Phát hiện lỗi xung đột port khi chạy inference, cần tối ưu cách khởi chạy server.

- Cấu trúc model trên Hugging Face chưa hoàn toàn phù hợp với Triton, cần điều chỉnh.

▼ 4. Upcomming

- Khắc phục lỗi xung đột port bằng cách cấu hình lại cổng trong Docker Compose và Script khởi chạy.
- Cải thiện cấu trúc model trên Hugging Face để phù hợp với định dạng yêu cầu của Triton Server.
- Đánh giá hiệu suất inference với cấu hình mới và tối ưu thêm nếu cần.
- Tiếp tục kiểm thử và push code lên GitHub sau khi hoàn thiện các điều chỉnh