Report Deploy Model

Report day: 18/02/2025

Name: Nguyễn Thanh Hòa

Team: 2 Team leader: Hòa

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

✅ **Task title 1: Deploy Triton Inference Server**

▼ 1. Purpose

Triển khai Triton Inference Server để phục vụ mô hình ONNX, gọi API inference và kiểm tra hiệu suất.

▼ 2. Action

- Tải image Triton Server từ NGC.
- Tạo repository chứa mô hình.
- Chạy Triton Server với model repository.
- Kiểm tra bằng cách gửi request inference.
- Push code lên GitHub.

▼ 3. Result

- Đã tải và chạy Triton Server thành công.
- Mô hình được host và có thể nhận request.
- Gọi API inference thành công với kết quả hợp lệ.
- Code đã được push lên GitHub.

▼ 4. Upcomming

Kiểm tra lại các config tối ưu hiệu suất.

Điều chỉnh tham số cấu hình để tối ưu tốc độ xử lý.

✅ **Task title 2: Performance Benchmarking**

▼ 1. Purpose

Đánh giá hiệu suất mô hình với Triton Performance Analyzer.

▼ 2. Action

- Cấu hình Perf Analyzer để đo lường hiệu suất.
- Chạy thử nghiệm với các chế độ tải khác nhau (Concurrency, Request Rate, Custom Interval).
- So sánh kết quả giữa các cấu hình.

▼ 3. Result

- Perf Analyzer chạy thành công.
- Đã thu thập dữ liệu về latency, throughput và utilization.
- Ghi nhận kết quả để cải thiện cấu hình Triton.

▼ 4. Upcomming

- Tối ưu model repository và server config.
- Tiếp tục phân tích hiệu suất và điều chỉnh thông số phù hợp.
- Đưa ra báo cáo tổng hợp về hiệu suất.