Nghiên cứu NAS, Ứng dụng, và Nén Đa tầng để Triển khai trên Edge

1. Mục tiêu đề tài

Nghiên cứu và tìm hiểu một phương pháp Tìm kiếm Kiến trúc Neron (NAS) cụ thể từ một bài báo khoa học (paper) và mã nguồn được cung cấp. Chạy thành công mã nguồn NAS để tạo ra một mô hình học sâu tối ưu cho một tác vụ. Cuối cùng, áp dụng một quy trình nén đa tầng (ví dụ: kết hợp tỉa thưa và lượng tử hóa) để tối ưu hóa mô hình đó và triển khai, đo hiệu xuất trên thiết bị biên (Raspberry Pi hoặc CPU Laptop).

2. Giai đoạn 1: Nghiên cứu và Ứng dụng NAS (NAS Research & Application)

- **Mục tiêu:** Hiểu và chạy thành công một công cụ NAS có sẵn.
- **Nhiệm vụ:**
    1. **Nghiên cứu Bài báo:** Học viên nhận một bài báo và link GitHub của một phương pháp NAS (theo link đã cung cấp). Học viên phải đọc, hiểu và viết một bản tóm tắt (1-2 trang) giải thích:
        - Không gian tìm kiếm (Search Space): Các khối xây dựng (building blocks) mà NAS có thể chọn là gì?
        - Chiến lược tìm kiếm (Search Strategy): NAS tìm kiếm kiến trúc tốt nhất như thế nào (ví dụ: tiến hóa, ngẫu nhiên, gradient-based)?
    2. **Chạy Mã nguồn NAS:** Cài đặt và chạy thành công mã nguồn NAS trên một bộ dữ liệu chuẩn (ví dụ: CIFAR-10 hoặc một bộ dữ liệu tùy chỉnh nhỏ) để *tạo ra* một kiến trúc mô hình mới.
    3. **Huấn luyện Baseline (Baseline Training):** Huấn luyện kiến trúc vừa được tìm thấy này từ đầu (from scratch) để có được độ chính xác cơ sở.

3. Giai đoạn 2: Quy trình Nén Mô hình Đa tầng (Multi-Stage Compression Pipeline)

- **Mục tiêu:** Tối đa hóa hiệu suất bằng cách *kết hợp* nhiều kỹ thuật nén.
- **Nhiệm vụ:** Học viên phải áp dụng một chuỗi các tối ưu hóa cho mô hình Model_NAS:
    1. **Tỉa thưa (Pruning)**
    2. **Tinh chỉnh (Fine-tuning):**
    3. **Lượng tử hóa (Quantization)**
    4. **.....**

4. Giai đoạn 3: Đo hiệu xuất và Phân tích Tổng hợp (Benchmarking & Analysis)

- **Mục tiêu:** Đánh giá hiệu xuất trên phần cứng mục tiêu.
- **Nhiệm vụ:**
    1. **Thiết lập Đo lường:** Xác định phần cứng mục tiêu (ví dụ: Raspberry Pi 5 hoặc CPU Laptop cụ thể) và các chỉ số đo lường (Độ chính xác, Độ trễ (ms), Kích thước

mô hình (MB)).
2. **Đo hiệu xuất của mô hình đã triển khai**.
3. **Phân tích:** Viết báo cáo phân tích xem NAS đã tạo ra mô hình tốt hơn không? Và quan trọng hơn, quy trình nén đa tầng (tỉa thưa + lượng tử hóa,.....) đã cải thiện *thêm* bao nhiêu.

🛠️ Sản phẩm chung cần nộp

1. **Báo cáo Cuối kỳ:** Một bài báo 6-8 trang bao gồm:
    - Tóm tắt phương pháp NAS đã nghiên cứu .
    - Mô tả quy trình nén đa tầng (Giai đoạn 2).
    - Trình bày và phân tích kết quả đo (Giai đoạn 3).
2. **Video Demo:** Một video 5 phút quay cảnh mô hình *cuối cùng* (Model_NAS_Final_INT8) chạy *thời gian thực* trên phần cứng mục tiêu và giải thích kết quả.
3. **Bài thuyết trình:** Thuyết trình 15 phút trước lớp.