Đề thi cuối kỳ

Môn: Deep Learning

**Bài 1:** Tìm hiểu chuyên sâu về các cơ chế Attention trong các mô hình ngôn ngữ lớn (LLMs)  
Chọn một nhóm cơ chế attention hiện đại (càng nhiều càng tốt, ít nhất là 4), trong danh sách dưới đây:

* 1. Self-Attention
  2. Multi-Query Attention (MQA)
  3. Grouped-Query Attention (GQA)
  4. FlashAttention hoặc FlashAttention v2
  5. Linear Attention
  6. Sparse Attention
  7. Rotary Positional Embedding (RoPE)

1. Trình bày lý thuyết chuyên sâu:
   * Giải thích cơ chế hoạt động, ý tưởng chính
   * Độ phức tạp tính toán
   * Ưu điểm và hạn chế
   * So sánh với self-attention gốc trong Transformer
   * Ứng dụng trong các mô hình LLM (large language model) hiện đại
2. Cài đặt mã nguồn minh họa (bằng PyTorch hoặc TensorFlow):
   * Viết lại cơ chế attention đã chọn dưới dạng module riêng
   * Tạo ví dụ đầu vào đơn giản (ví dụ: chuỗi ngắn hoặc dãy số)
   * In ra và giải thích ma trận attention tương ứng
3. Nộp bài:
   * Bài viết trình bày lý thuyết (PDF hoặc Word)
   * Notebook minh họa (Colab hoặc Jupyter)

**Bài 2:**Xây dựng mô hình trích xuất văn bản từ ảnh bằng mô hình kết hợp CNN và Transformer-Decoder  
**Yêu cầu:**

1. Tìm hiểu và trình bày nguyên lý hoạt động của kiến trúc kết hợp CNN + Transformer Decoder trong bài toán trích xuất văn bản từ ảnh (image-to-text OCR):
   * CNN dùng để trích đặc trưng hình ảnh
   * Transformer Decoder dùng để sinh chuỗi văn bản
   * Cơ chế attention giữa đặc trưng ảnh và trình giải mã văn bản
2. Triển khai mô hình đơn giản để trích xuất văn bản từ ảnh:
   * Sử dụng tập ảnh có chứa văn bản (ảnh quảng cáo, ảnh thẻ, hóa đơn, biển hiệu, v.v.)
   * Áp dụng CNN backbone (ResNet hoặc đơn giản hơn)
   * Kết hợp Transformer Decoder để sinh ra văn bản tương ứng
3. Có thể sử dụng thư viện hỗ trợ như PyTorch hoặc TensorFlow, nhưng không được dùng các mô hình OCR có sẵn như EasyOCR, TrOCR, v.v.
4. Đánh giá độ chính xác trên tập Test
5. Hiển thị kết quả đầu ra:
   * Ảnh gốc
   * Văn bản được trích xuất
   * Nếu có thể, vẽ heatmap attention để minh họa
6. Gửi kèm:
   * File báo cáo trình bày nguyên lý mô hình
   * Notebook hoặc Colab mô hình chạy được

**Lưu ý:**

1. Tất cả các thành viên trong nhóm đều phải nộp bài;
2. Nộp báo cáo Word hoặc Pdf, các file code, còn data thì chỉ gửi link
3. Code nộp thêm link colab chạy được