1. Lý thuyết chung và mô hình sử dụng:
   * Lý thuyết chung: Trích xuất dữ liệu từ các nguồn như văn bản, ảnh, âm thanh, video là một bài toán rất quan trọng trong xử lý ngôn ngữ tự nhiên (NLP) và thị giác máy tính (CV). Mô hình Transformer based Encoder-Decoder đã được chứng minh là rất hiệu quả trong các bài toán này.
   * Mô hình sử dụng: Đề xuất sử dụng mô hình Transformer based Encoder-Decoder như BERT (Bidirectional Encoder Representations from Transformers) kết hợp với một Decoder dựa trên cơ chế Attention như Seq2Seq hoặc Transformer Decoder. Điều này sẽ cho phép mô hình hiểu ngữ cảnh và đặc trưng trong dữ liệu đầu vào, đồng thời sinh ra dữ liệu đầu ra một cách có cấu trúc và liền mạch.
2. Xây dựng dữ liệu:
   * Ví dụ về bài toán mô tả món ăn từ ảnh menu:
     + Thu thập dữ liệu ảnh menu và mô tả văn bản tương ứng.
     + Xử lý ảnh bằng các kỹ thuật như cắt, chuẩn hóa kích thước, ...
     + Xử lý văn bản bằng các kỹ thuật như tokenization, padding, ...
3. Huấn luyện mô hình:
   * Chia tập dữ liệu thành tập huấn luyện, tập validation và tập test.
   * Sử dụng mô hình Transformer based Encoder-Decoder, khởi tạo trọng số từ các mô hình tiền huấn luyện như BERT.
   * Tiến hành huấn luyện mô hình bằng cách tối ưu hóa hàm mất mát phù hợp như cross-entropy.
4. Test và đánh giá:
   * Sử dụng tập test để đánh giá độ chính xác của mô hình.
   * Các chỉ số đánh giá có thể bao gồm: BLEU score, METEOR, ROUGE, ... tùy thuộc vào bài toán.
   * Đánh giá kết quả, phân tích lỗi và cải thiện mô hình nếu cần thiết.