OBJECT RECOGNITION AS NEXT TOKEN PREDICTION

Nguyễn Xuân Bách

Trường ĐH Công Nghệ Thông Tin-UIT

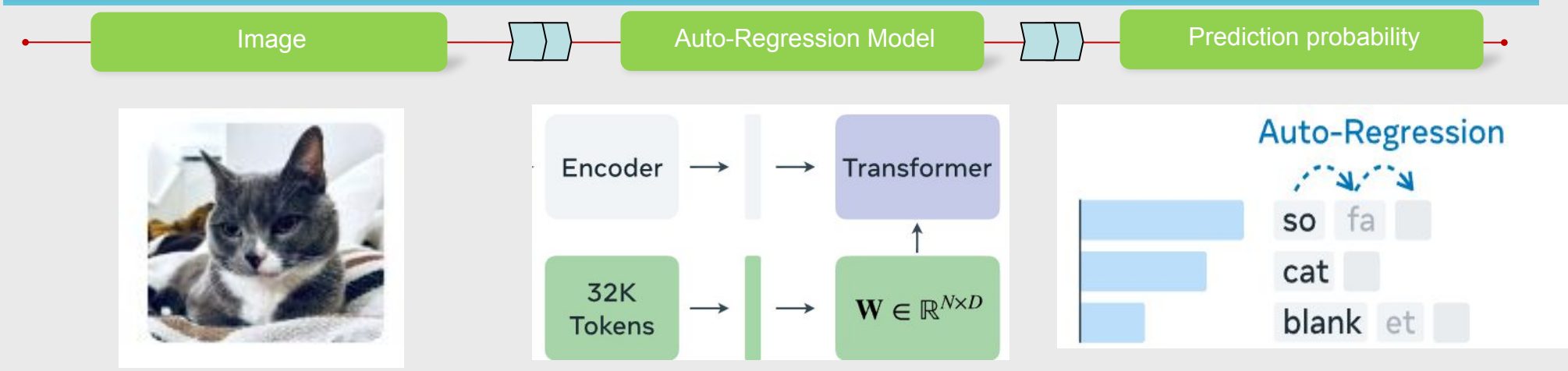
What ?

Đề tài "Object Recognition as Next Token Prediction" nghiên cứu cách nhận dạng đối tượng trong thị giác máy tính bằng cách sử dụng mô hình ngôn ngữ lớn (LLM) và mô hình hình ảnh CLIP. Phương pháp này loại bỏ nhu cầu định trước nhãn đối tượng, thay vào đó, sử dụng vectơ nhúng từ mô tả đầu vào để dự đoán nhãn từ hình ảnh.

Why ?

Chúng tôi chọn đề tài "Object Recognition as Next Token Prediction" vì nó mang lại một phương pháp mới mẻ và linh hoạt trong nhận dạng đối tượng, vượt qua hạn chế của việc định trước nhãn đối tượng. Bằng cách tích hợp mô hình ngôn ngữ lớn (LLM) và CLIP, phương pháp này hứa hẹn nâng cao khả năng nhận diện đối tượng trong các tình huống thực tế đa dạng và phức tạp.

Overview



Description

1. Aligning Images and Text

Hình ảnh 2D được đưa vào mô hình nền tảng như ViT trong CLIP để tạo ra các vectơ nhúng hình ảnh Xv. Mục tiêu là giải mã các nhãn đối tượng từ Xv bằng cách căn chỉnh với các embedding văn bản W.

2. Auto-Regression for Recognition

Sử dụng các embedding token từ LLM được huấn luyện trước như LLaMA để định nghĩa W. Mô hình sử dụng bộ giải mã ngôn ngữ để dự đoán các token từ Xv bằng cách kết hợp chúng với các embedding văn bản W.

3. Non-causal Masking

Để dự đoán tất cả các nhãn đối tượng trong một hình ảnh, chúng tôi sử dụng mặt nạ không nhân quả để tách rời các token từ các nhãn khác nhau. Điều này cho phép các token của mỗi nhãn chỉ tham gia vào các token từ cùng một nhãn, giúp mô hình dự đoán các nhãn độc lập nhưng vẫn duy trì mối quan hệ không gian của các token hình ảnh.

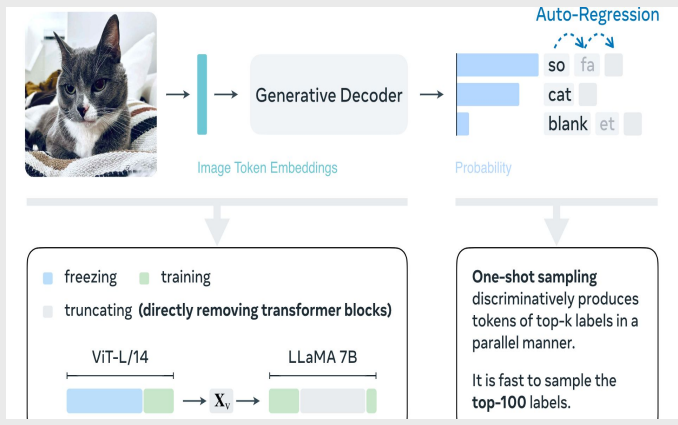
4. One-Shot Sampling

Chúng tôi áp dụng phương pháp lấy mẫu một lần để dự đoán đồng thời các token của nhiều nhãn đối tượng. Sau khi xác định các token bạn đầu có xác suất cao nhất từ đầu vào X, chúng tôi tiếp tục lấy mẫu các token tiếp theo cho các nhãn chưa hoàn thành, sử dụng top-1 sampling. Phương pháp này tối ưu hóa quá trình lấy mẫu bằng cách hoạt động song song và tránh vấn đề lặp lại trong các phương pháp tìm kiếm tham lam và tìm kiếm chùm.

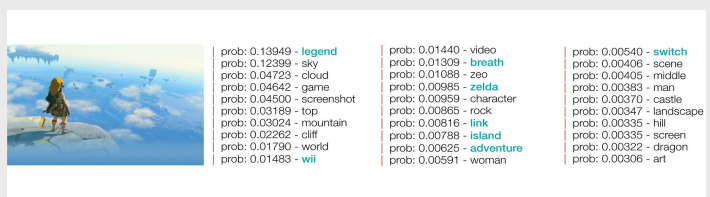
5. Truncating the Decoder

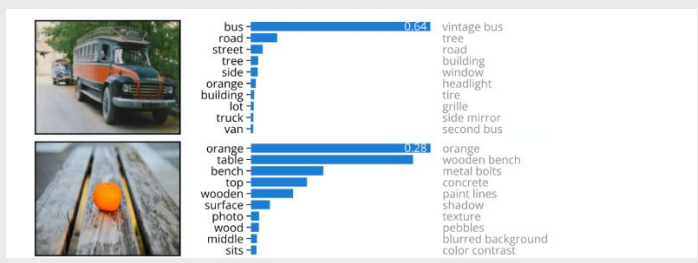
Chúng tôi cải thiện hiệu quả bằng cách cắt ngắn bộ giải mã ngôn ngữ LLaMA 7B, giữ lại 6 khối transformer đầu tiên và lớp đầu ra cuối cùng. Bộ giải mã ngôn ngữ cắt ngắn này, Langtruncated, vẫn duy trì bộ mã hóa từ và các embedding token đã được huấn luyện trước để mã hóa đầu vào, giúp tối ưu hóa hiệu suất mà không ảnh hưởng đến độ chính xác.

Pipeline



Example result





NII

Nguyễn Xuân Bách– UIT

TEL :0904820419 Email : 22520093@gm.uit.edu.vn