NHẬN DẠNG ĐỐI TƯỢNG QUA DỰ ĐOÁN MÃ THÔNG BÁO TIẾP THEO

Nguyễn Xuân Bách - 22520093

Tóm tắt

- Lớp: CS519.O21.KHTN
- Link Github: https://github.com/bach04py/CS519.O21.KHTN.git
- Link YouTube video: https://youtu.be/l-zokgu3cls
- Ảnh + Họ và Tên:Nguyễn Xuân Bách
- 

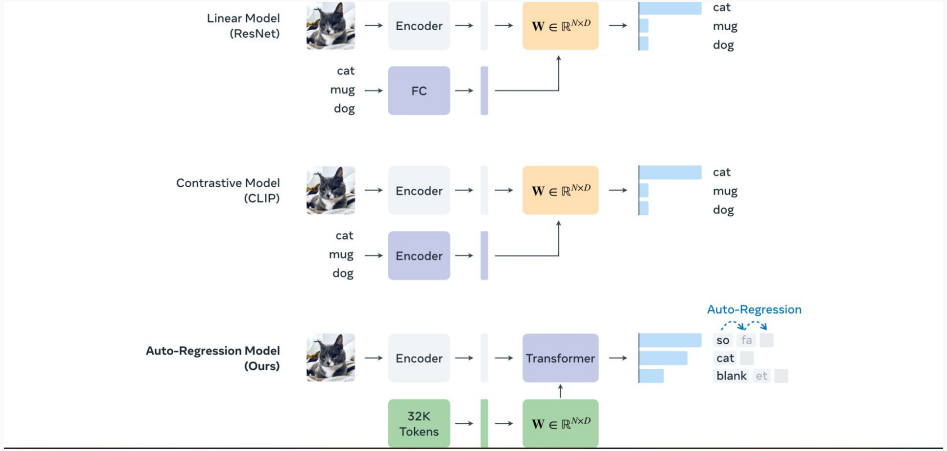
Giới thiệu

Đề tài "Object Recognition as Next Token Prediction" tập trung vào việc giải quyết một vấn đề cơ bản trong thị giác máy tính - nhận dạng đối tượng. Chúng tôi đề xuất một cách tiếp cận mới, sử dụng mô hình ngôn ngữ lớn (LLM) và mô hình hình ảnh CLIP để nhận dạng đối tượng mà không cần định trước nhãn đối tượng. Điều này được thực hiện thông qua việc sử dụng mô hình ngôn ngữ để tạo ra các vectơ nhúng từ các mô tả đầu vào và sử dụng các vectơ nhúng này để dự đoán các nhãn đối tượng từ hình ảnh.

Figure 1. **Object recognition as next token prediction** using a generative decoder such as a transformer-based language model to auto-regressively predict object labels. *Photo authorized with CC BY 4.0.*

Mục tiêu

- Xây dựng thuật toán nhận dạng đối tượng từ hình ảnh dựa trên dự đoán mã thông báo tiếp theo và so sánh độ chính xác, tốc độ xử lý, và khả năng tổng quát hóa với các phương pháp khác.
- Đánh giá hiệu suất thuật toán trên các bộ dữ liệu đa dạng, bao gồm đối tượng phổ biến và hiếm, bằng cách đo độ chính xác, độ phủ, và khả năng xử lý đối tượng hiếm.
- Tối ưu hóa thuật toán để đạt hiệu suất tốt với tài nguyên tính toán hạn chế, đo lường thông qua thời gian huấn luyện, kích thước mô hình, và hiệu suất nhận dạng đối tượng.



Nội dung và Phương pháp

Chúng tôi tập trung vào việc nhận dạng đối tượng trong hình ảnh bằng cách sử dụng mô hình ngôn ngữ lớn (LLM) và mô hình CLIP, mà không cần thông tin về các nhãn đối tượng đã được định trước. Phương pháp đề xuất sử dụng các embedding văn bản từ LLM để tạo ra các trọng số động cho việc nhận dạng, thay vì sử dụng các trọng số tĩnh cố định như trong các mô hình truyền thống.

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.

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.

Nội dung và Phương pháp

Non-casual Masking:

Để dự đoán tất cả các nhãn đối tượng trong một hình ảnh, chúng tôi sử dụng mặt nạ non -casual 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.

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 ban 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.

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.

Kết quả dự kiến

- · **Khả năng Nhận dạng Đối tượng Hiệu quả: Thuật toán đề xuất có khả năng nhận dạng các đối tượng từ hình ảnh một cách hiệu quả và chính xác. Điều này được đánh giá thông qua các thí nghiệm so sánh với các phương pháp nhận dạng đối tượng hiện tại, như mạng nơ-ron tích chập (CNN).**
- · **Hiệu suất trên bộ dữ liệu đa Dạng: Thuật toán được thử nghiệm trên nhiều bộ dữ liệu khác nhau, bao gồm cả những đối tượng phổ biến và hiếm, và đạt được độ chính xác cao, khả năng phủ rộng và khả năng xử lý tốt các đối tượng hiếm.**
- · **Tối ưu Hóa Tài Nguyên: Thuật toán được tối ưu hóa để hoạt động hiệu quả với tài nguyên tính toán hạn chế, chẳng hạn như giảm thời gian huấn luyện và kích thước mô hình, trong khi vẫn duy trì hiệu suất cao trong nhận dạng đối tượng.**

Tài liệu tham khảo

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur

Mensch, Katherine Millican, Malcolm Reynolds, et al.

Flamingo: A Visual Language Model for Few-Shot Learning. In NeurIPS, 2022. 1, 2, 13

[2] Jacob Andreas and Dan Klein. Reasoning About Pragmatics With Neural Listeners and Speakers. In EMNLP, 2016.2

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel,

Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan

Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo:

An Open-Source Framework for Training Large Autoregressive Vision-Language Models. In arXiv:2308.01390,

2023. 1, 5, 6, 9, 12, 13

Tài liệu tham khảo

[4] Kobus Barnard and David Forsyth. Learning the Semantics

of Words and Pictures. In ICCV, 2001. 14

[5] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando

De Freitas, David M Blei, and Michael I Jordan. Matching Words and Pictures. In JMLR, 2003. 14

[6] Abhijit Bendale and Terrance Boult. Towards Open World

Recognition. In CVPR, 2015. 2

[7] Rodrigo Benenson and Vittorio Ferrari. From Colouring-in

to Pointillism: Revisiting Semantic Segmentation Supervision. arXiv:2210.14142, 2022. 4

[8] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing With Python: Analyzing Text With the

Natural Language Toolkit. O'Reilly Media, Inc., 2009. 12