BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS519 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS519.N11

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

UIT-HCM

TÊN ĐỀ TÀI - NHẬN DIỆN VẬT THỂ VỚI MÔ HÌNH TRANSFORMERS

Mai Hiếu Hiền - 20521305
Nguyễn Hoàng Hải - 20521279
Nguyễn Thị Kim Anh - 20521072

Tóm tắt

- Link Github của nhóm: https://github.com/maihieuhien/CS519.N11
- Link YouTube video: https://youtu.be/vAQ_YNCVKHE
- Ảnh + Họ và Tên của các thành viên

Mai Hiếu Hiền

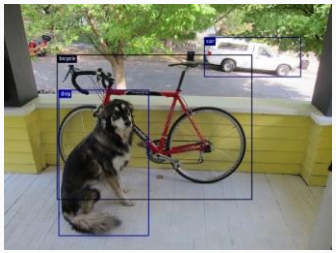
Nguyễn Hoàng Hải

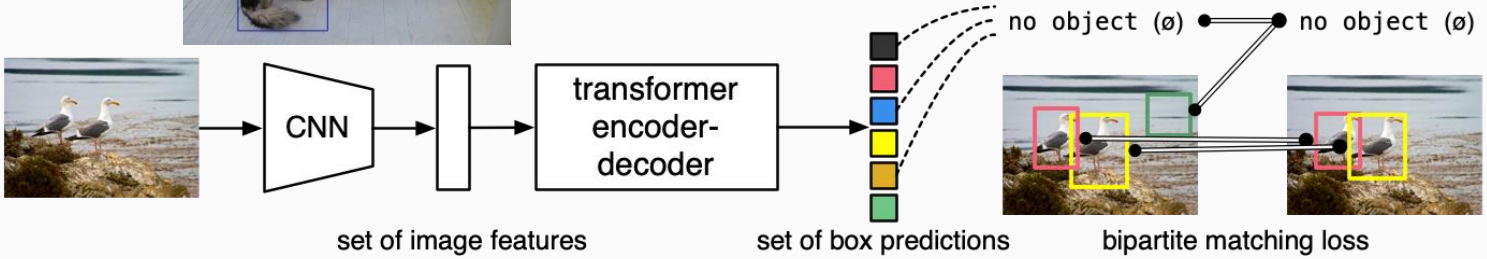
Nguyễn Thị Kim Anh

Giới thiệu

- Input:

- Output: Lớp của vật thể và vị trí của vật thể





Giới thiệu

- Sự đột phá của các kiến trúc CNN và học sâu đã cải thiện hiệu quả của các mô hình nhận dạng rất nhiều so với các phương pháp mô tả đặc trưng.
- Các mô hình hiện đại như Faster-RCNN [3], SSD [4], RetinaNet [5] sử dụng Anchor boxes để dự đoán các bounding boxes của các vật thể và kết hợp Non-maximum suppression (NMS) để cải thiện hiệu suất các mô hình.
- Các mô hình dựa trên anchor rất khó khi áp dụng cho các dữ liệu khác nhau và yêu cầu người nghiên cứu phải có kinh nghiệm. Nhiều mô hình nhận dạng không dựa trên anchor (anchor-free) ra đời như CornerNet [6], CenterNet [7] đã giải quyết được vấn đề trên.

Mục tiêu

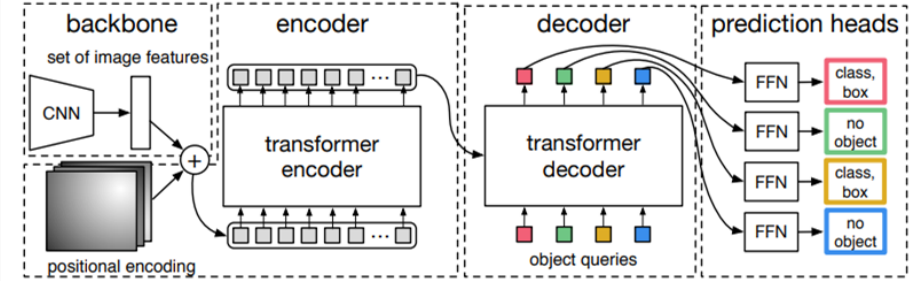
- Đề xuất các mô hình Transformers cho bài toán nhận diện vật thể có hiệu quả tốt hơn mô hình anchor-based và anchor-free trước đó trên độ đo AP trên bộ dữ liệu COCO 2017.
- Xây dựng một trang web demo để trực quan hóa hiệu quả của mô hình Transformers cho bài toán nhận diện vật thể.

Nội dung và Phương pháp

Nội dung:

- **"Làm sao để mô hình nhận dạng không sử dụng NMS (NMS để loại bỏ các bounding box không cần thiết) ? tức là các bounding box phải có mối liên hệ với nhau để tránh trùng lặp"**, điều này rất giống với cơ chế Attention (**Attention mechanism**). Từ giả thuyết này, chúng tôi khảo sát các kiến trúc Transformers khác nhau với cơ chế Attention [8] đã được đề xuất trước đó.

(Hình 1)



- Khảo sát về độ hiệu quả và các kĩ thuật của các mô hình anchor-based [3][4][5] và anchor-free [6][7] trước đó. Từ đó sẽ đề xuất hàm lỗi và một số kĩ thuật để tăng hiệu quả mô hình Transformers.

- Khảo sát cách matching giữa bounding boxes từ output của mô hình với ground truth data theo các mô hình đã đề xuất trước đó [9].

Nội dung và Phương pháp

Phương pháp:

- Sẽ tiến hành khảo sát các bài báo liên quan tại các hội nghị CVPR, ICCV, ECCV, NIPS, ICLR.
- Nghiên cứu về việc sử dụng Backbone [9] với các mô hình được huấn luyện sẵn khác nhau (VGG [10], Resnet [11]) – một phương pháp phổ biến hiện nay – với Positional Encoding khác nhau (spatial positional encoding và output positional encodings) của mô hình Transformers [8].
- Từ output của transformer decoder, chúng tôi đề xuất sử dụng mạng feed-forward networks (FFN) để dự đoán các bounding box và lớp của các bounding box.
- Nghiên cứu thuật toán Hungarian [13] để matching giữa bounding boxes từ mô hình với ground truth data và thiết kế hàm lỗi cho mô hình Transformers.
- Các mô hình Transformers được đề xuất sẽ được huấn luyện và đánh giá dựa trên bộ data COCO 2017 [12] và so sánh với các mô hình khác.

Kết quả dự kiến

- Mô hình Transformers đã được huấn luyện trên bộ dữ liệu COCO 2017 phải có kết quả theo độ đo AP cao hơn Faster-RCNN [3], SSD [4], RetinaNet [5], CornerNet [6], CenterNet [7]
- Một bài báo tại hội nghị quốc tế
- Một trang web demo để giới thiệu và minh họa cho nghiên cứu

Tài liệu tham khảo

[1]. Navneet Dalal, Bill Triggs:
Histograms of Oriented Gradients for Human Detection. CVPR (1) 2005: 886-893
[2]. Paul A. Viola, Michael J. Jones:
Rapid Object Detection using a Boosted Cascade of Simple Features. CVPR (1) 2001: 511-518
[3]. Shaoqing Ren, Kaiming He, Ross B. Girshick, Jian Sun:
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS 2015: 91-99
[4]. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, Alexander C. Berg:
SSD: Single Shot MultiBox Detector. ECCV (1) 2016: 21-37
[5]. Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, Piotr Dollár:
Focal Loss for Dense Object Detection. ICCV 2017: 2999-3007
[6]. Hei Law, Jia Deng:
CornerNet: Detecting Objects as Paired Keypoints. ECCV (14) 2018: 765-781
[7]. Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, Qi Tian:
CenterNet: Keypoint Triplets for Object Detection. ICCV 2019: 6568-6577
[8]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin:
Attention is All you Need. NIPS 2017: 5998-6008
[9]. Russell Stewart, Mykhaylo Andriluka, Andrew Y. Ng:
End-to-End People Detection in Crowded Scenes. CVPR 2016: 2325-2333