

BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học

**CS519 - PHƯƠNG PHÁP LUẬN
NGHIÊN CỨU KHOA HỌC**

Lớp học

CS519.N11

Giảng viên

PGS.TS. LÊ ĐÌNH DUY


Thời gian

09/2022 - 02/2023

----- *Trang này cố tình để trống* -----

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
https://youtu.be/vAQ_YNCVKHE
- Link slides (dạng .pdf đặt trên Github của nhóm):
<https://github.com/maihieuhien/CS519.N11/blob/main/slides.pdf>
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">● Họ và Tên: Mai Hiếu Hiền● MSSV: 20521305 	<ul style="list-style-type: none">● Lớp: CS519.N11● Tự đánh giá (điểm tổng kết môn): 9.5/10● Số buổi vắng: 0● Số câu hỏi QT cá nhân: 11● Số câu hỏi QT của cả nhóm: 3● Link Github: https://github.com/maihieuhien/CS519.N11/● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng nghiên cứu○ Làm slide thuyết trình và chỉnh sửa slide○ Lên ý tưởng và làm video YouTube
<ul style="list-style-type: none">● Họ và tên: Nguyễn Hoàng Hải● MSSV: 20521279	<ul style="list-style-type: none">● Lớp CS519.N11● Tự đánh giá (điểm tổng kết môn): 9/10● Số buổi vắng: 3● Số câu hỏi QT cá nhân: 1● Số câu hỏi QT của cả nhóm: 3● Link Github:

	<p>https://github.com/maihieuhien/CS519.N11/</p> <ul style="list-style-type: none"> ● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: <ul style="list-style-type: none"> ○ Làm slide thuyết trình và chỉnh sửa slide ○ Lên ý tưởng và làm video YouTube
<ul style="list-style-type: none"> ● Họ và tên: Nguyễn Thị Kim Anh ● MSSV: 20521072 	<ul style="list-style-type: none"> ● Lớp: CS519.N11 ● Tự đánh giá (điểm tổng kết môn): 8.8/10 ● Số buổi vắng: 3 ● Số câu hỏi QT cá nhân: 10 ● Số câu hỏi QT của cả nhóm: 3 ● Link Github: <p>https://github.com/maihieuhien/CS519.N11/</p> ● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: <ul style="list-style-type: none"> ○ Làm slide thuyết trình và chỉnh sửa slide ○ Lên ý tưởng và làm video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

NHẬN DIỆN VẬT THỂ VỚI MÔ HÌNH TRANSFORMERS

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

OBEJECT DETECTION WITH TRANSFORMERS MODEL

TÓM TẮT *(Tối đa 400 từ)*

Hiện nay, các mô hình học sâu học có kiến trúc convolutional neural network (CNN) được sử dụng phổ biến để giải quyết bài toán nhận diện vật thể (object detection). Tuy nhiên, các mô hình CNN kết hợp các thành phần được thiết kế bởi con người như a non-maximum suppression hoặc anchor generation yêu cầu rất nhiều kinh nghiệm trong nghiên cứu để mô hình hiệu quả hơn so với các nghiên cứu trước. Với những vấn đề trên, nghiên cứu này đề xuất sử dụng mô hình Transformers để giải quyết bài toán nhận diện vật thể mà loại bỏ các thành phần được thiết kế bằng tay. Chúng tôi sẽ sử dụng bộ dữ liệu COCO 2017 và để huấn luyện, đánh giá và so sánh hiệu quả của mô hình Transformers với các mô hình CNN trước đó với độ đo phổ biến trong bài toán nhận diện vật thể là AP.

GIỚI THIỆU *(Tối đa 1 trang A4)*

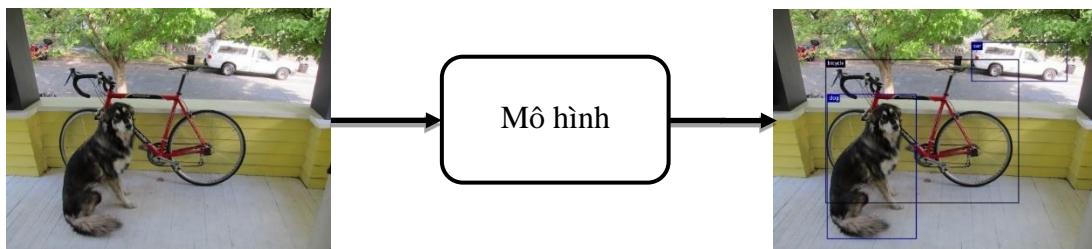
Nhận diện vật thể là một trong những bài toán quen thuộc của lĩnh vực thị giác máy tính mà phân loại và nhận dạng các vật thể trong một video hoặc ảnh. Một số mô hình nhận dạng đầu tiên sử dụng các phương pháp mô tả đặc trưng như Histogram of Oriented Gradients (HOG) [1], Viola-Jones detector [2]. Sự 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. Mặc dù phương pháp Anchor boxes có kết quả tốt so với các phương pháp trước đó, việc cài đặt 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. Do đó, có 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. Do đó trong nghiên cứu này, chúng tôi nghiên cứu và đề xuất sử dụng kiến trúc Transformers [8], mô hình anchor-free, dựa trên các nhược điểm của các mô hình anchor-based và mong muốn hiệu quả cao hơn so với các mô hình anchor-based và anchor-free trước đó theo độ đo Average Precision trên bộ dữ liệu COCO [12].

Input: Một bức ảnh

Output: Vị trí của các vật thể và lớp của vật thể đó

Ảnh minh họa:



MỤC TIÊU (*Viết trong vòng 3 mục tiêu*)

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

NỘI DUNG VÀ PHƯƠNG PHÁP

Để xây dựng được mô hình Transformers theo mục tiêu đề ra, chúng tôi đề xuất các nội dung nghiên cứu sau:

Nội dung:

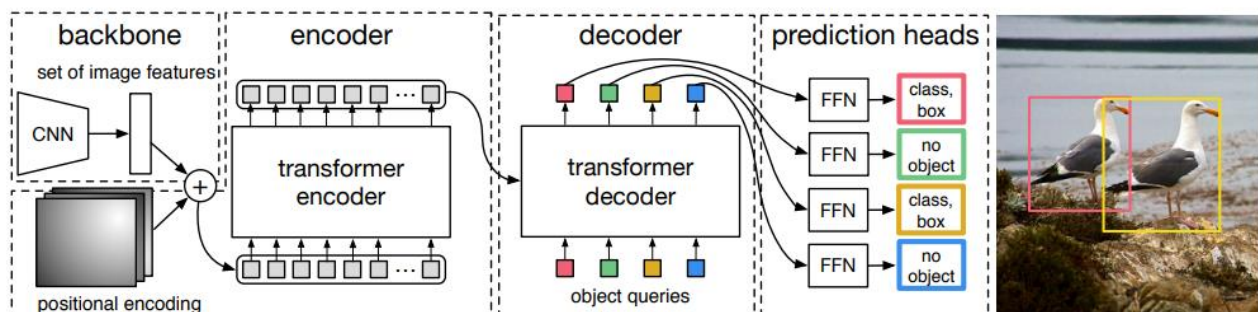
- Đặt giả thiết: output của một mô hình nhận dạng anchor-based sẽ bao gồm rất nhiều bounding boxes và vì vậy sẽ có rất nhiều bounding box cho một đối tượng (object) và sẽ dẫn đến dư thừa. Do đó phải dùng NMS để loại bỏ các bounding box không cần thiết. Vậy **“Làm sao để mô hình nhận dạng không**

sử dụng NMS ? 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 thiết này, chúng tôi khảo sát các kiến trúc Transformers khác nhau với cơ chế Attention đã được đề xuất trước [8]

- Khảo sát về độ hiệu quả và các kỹ thuật của các mô hình anchor-based [3][4][5] và anchor-free [6][7] trước đó. Từ đó sẽ đề xuất hàm lỗi (loss function) và một số kỹ 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.

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.
- Mô hình Transformers cho bài toán nhận diện được đề xuất theo hình 1:



Hình 1: Mô hình Transformers cho bài toán nhận diện

- Khảo sát và nghiên cứu thuật toán Hungarian [13][14] để matching giữa bounding boxes từ mô hình với ground truth 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 MONG ĐỢI

- Mô hình Transformers đã được huấn luyện trên bộ dữ liệu COCO 2017 [12] 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 *(Định dạng DBLP)*

[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]. Luis Pineda, Amaia Salvador, Michal Drozdal, Adriana Romero:

Elucidating image-to-set prediction: An analysis of models, losses and datasets. CoRR abs/1904.05709 (2019)

[10]. Karen Simonyan, Andrew Zisserman:

Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015

[11]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun:

Deep Residual Learning for Image Recognition. CVPR 2016: 770-778

[12]. Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick:

Microsoft COCO: Common Objects in Context. ECCV (5) 2014: 740-755

[13]. Russell Stewart, Mykhaylo Andriluka, Andrew Y. Ng:

End-to-End People Detection in Crowded Scenes. CVPR 2016: 2325-2333

[14]. Jesmin Jahan Tithi, Sriram Aananthakrishnan, Fabrizio Petrini:

Online and Real-time Object Tracking Algorithm with Extremely Small Matrices. CoRR abs/2003.12091 (2020)