THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://www.youtube.com/channel/UCvpoCxCfSWmzeNyJ6HEzu-A
- Link slides (dạng .pdf đặt trên Github của nhóm):
  https://github.com/Kryst4lize/CS519.O11/blob/BTVN/il_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*

• Họ và Tên: Lê Thanh Minh • MSSV: 21520063 	• Lớp: CS519.O11 • Tự đánh giá (điểm tổng kết môn): 8.5/10 • Số buổi vắng: 1 • Số câu hỏi QT cá nhân: 3 • Số câu hỏi QT của cả nhóm: 3 • Link Github: https://github.com/mynameuit/CS519.O11/ • Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: ○ Lên ý tưởng đồ án ○ Viết đề cương, làm slide thuyết trình ○ Làm video YouTube, Poster

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

MÔ HÌNH HỌC TIỆM TIẾN ĐỂ KHAI THÁC TẬP DỮ LIỆU NGOÀI TRONG BÀI TOÁN NHẬN DIỆN VẬT THỂ

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

INCREMENTAL LEARNING MODEL FOR OBJECT DETECTION: LEVERAGING EXTERNAL DATA FOR IMPROVEMENT

TÓM TẮT *(Tối đa 400 từ)*

Trong lĩnh vực thị giác máy tính và trí tuệ nhân tạo, việc nhận diện vật thể đang là một trong những đề tài quan trọng. Tuy nhiên, có hai thách thức chưa được chú ý đúng mức. Thứ nhất, các nghiên cứu chủ yếu tập trung vào cải thiện độ chính xác và tốc độ xử lý dữ liệu trên các tập dữ liệu như CIFAR-100 và ImageNet100, nhưng chưa tập trung đến việc xây dựng mô hình có khả năng huấn luyện trên dữ liệu lớn hoặc sử dụng nhiều tập dữ liệu. Thứ hai, tập dữ liệu hiện tại chỉ tập trung vào các đối tượng được gắn nhãn, bỏ qua các đối tượng khác trong ảnh, không phản ánh thực tế khi số đối tượng thực tế nhiều hơn rất nhiều. Nghiên cứu này nhấn mạnh việc xây dựng mô hình có tính mở rộng, có khả năng sử dụng dữ liệu bên ngoài tập dữ liệu có sẵn. Điều này được đạt được bằng cách cập nhật mô hình khi có dữ liệu mới, giúp tận dụng kho dữ liệu ngày càng lớn từ Internet và nhận diện nhiều vật thể hơn. Các phương pháp tiếp cận trước đây thường sử dụng thêm tham số để lưu trữ đặc trưng của các lớp đối tượng mới, tăng số lượng tham số khi dữ liệu gia tăng. Đề xuất sử dụng mô hình Transformer giúp biểu diễn các lớp đối tượng cũ và mới với số lượng tham số như nhau, đảm bảo tính thống nhất khi dữ liệu tăng. Ngoài ra, để rút trích đặc trưng tốt nhất, mô hình sử dụng các mạng nơ-ron tích chập tiền huấn luyện như VGG-16, Resnet50, Inception. Đồng thời, bộ dữ liệu cũng được xử lý để phù hợp với bài toán nhận diện đối tượng đa dạng. Tổng cộng, nghiên cứu này đề xuất một hướng tiếp cận mới để giải quyết thách thức của việc nhận diện vật thể trong môi trường đa dạng và lớn lên ngày càng mở rộng.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Nhận diện vật thể đang là một ưu tiên quan trọng trong lĩnh vực thị giác máy tính và trí tuệ nhân tạo. Tuy nhiên, bài toán này đối mặt với hai thách thức chưa được đặc biệt chú ý. Thứ nhất, các nghiên cứu hiện tại chủ yếu tập trung vào việc cải thiện độ chính xác và tốc độ xử lý của mô hình trên các

tập dữ liệu như CIFAR-100 [1] và ImageNet100 [2], nhưng thiếu sự quan tâm đúng mức đến khả năng huấn luyện mô hình trên dữ liệu khổng lồ từ bên ngoài hoặc sử dụng nhiều tập dữ liệu. Thứ hai, tập dữ liệu hiện tại chỉ tập trung vào các đối tượng được gắn nhãn, bỏ qua các đối tượng xuất hiện trong ảnh mà không có nhãn, không phản ánh đầy đủ thực tế khi số lượng đối tượng thực tế có thể lớn hơn nhiều so với tập dữ liệu.

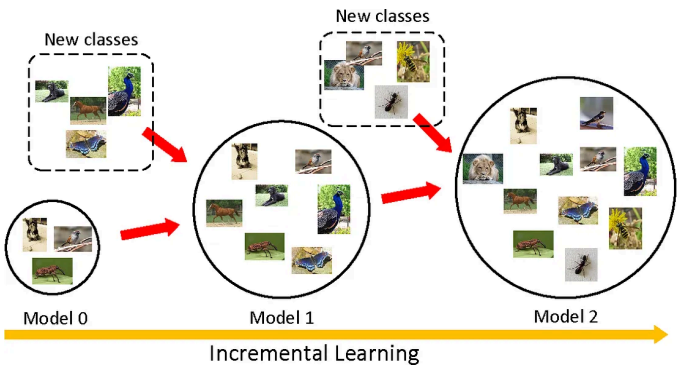
Điều này đặt ra thách thức về việc xây dựng mô hình có khả năng mở rộng (Scalability). Mô hình cần có khả năng tirn dụng dữ liệu bên ngoài tập dữ liệu có sẵn, cập nhật mô hình khi có dữ liệu mới, và đồng thời có khả năng nhận diện nhiều đối tượng hơn. Trong ngữ cảnh này, các phương pháp trước đây [3][4] đã sử dụng thêm tham số để lưu trữ thêm đặc trưng của các lớp đối tượng mới, tuy nhiên, điều này dẫn đến vấn đề khi số lượng tham số tăng lên theo độ lớn của dữ liệu khiến cho mô hình trở nên quá tải khi mở rộng quá nhiều đối tượng

Để giải quyết vấn đề này, đề xuất sử dụng mô hình Transformer [5]. Phương pháp này đảm bảo tính thống nhất bằng cách biểu diễn các lớp đối tượng cũ và mới với cùng một số lượng tham số, giúp mô hình duy trì hiệu suất khi có thêm dữ liệu mới. Bên cạnh đó, cần xử lý bộ dữ liệu để đảm bảo phù hợp với yêu cầu của bài toán. Để rút trích đặc trưng một cách hiệu quả, sẽ sử dụng các mô hình Pre-trained của họ hàng nhà CNN [9] như VGG-16 [6], Resnet50 [7], Inception [8].

Tóm lại, đề xuất của chúng tôi đó là xây dựng một mô hình tiệm tiến với kiến trúc Transformer, giúp mô hình có khả năng học và nhận diện các lớp mới và lớp đối tượng cũ trên cả dữ liệu cụ thể và dữ liệu mới mở rộng.

Input : Tập dữ liệu ban đầu (Base) và các tập dữ liệu bổ sung (Extra) theo thời gian (Dữ liệu bổ sung bổ sung số lượng lớp đối tượng mới)

Output: Mô hình có thể học nhận diện các lớp mới mới và lớp đối tượng cũ



MỤC TIÊU

- Thu thập dữ liệu và xây dựng một bộ dữ liệu UIT_dataset phù hợp với việc đánh giá mô hình học tiệm tiến

- Xây dựng mô hình tiệm tiến dựa trên kiến trúc Transformer
- Tinh chỉnh mô hình sao cho đạt hiệu suất cao nhất

NỘI DUNG VÀ PHƯƠNG PHÁP

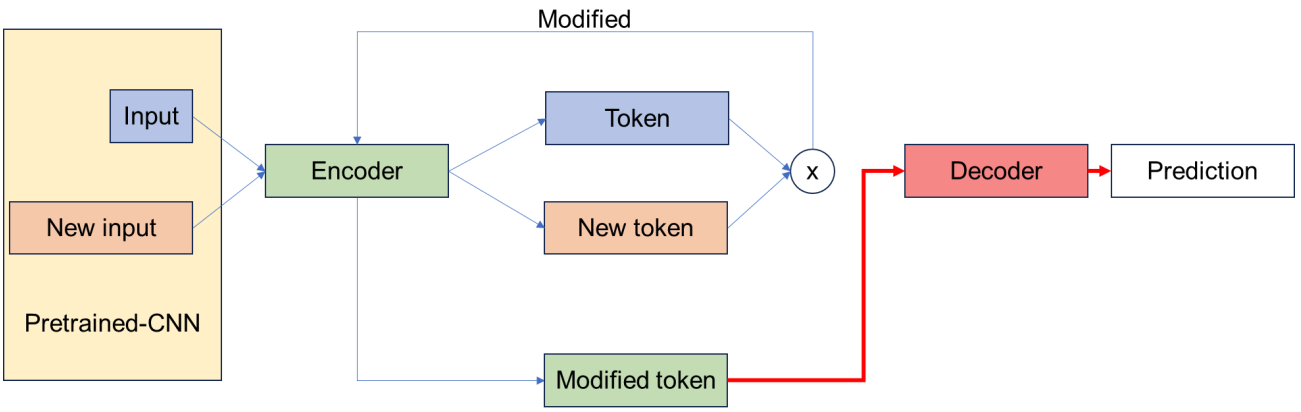
Nội dung 1: Thu thập dữ liệu và xây dựng bộ dữ liệu phù hợp cho mô hình học tiệm tiến.

- **Phương pháp thực hiện:**
  + Thu thập dữ liệu từ các nguồn uy tín, được sử dụng làm thước đo cho nhiều bài báo khoa học ở các hội nghị uy tín như ImageNet100, CIFAR-10, MS COCO [10],
  + Chia tập dữ liệu này thành hai phần : Base (chứa 50% lớp đối tượng của tập dữ liệu); Extra (Gồm 10 phần nhỏ (Batch), mỗi phần là tập dữ liệu chứa 5% số lớp đối tượng của tập dữ liệu đó). Các class trong phần Base và các Batch trong phần Extra không trùng nhau.

    Vd: ImagetNet100 có tổng cộng 100 class, được chia làm 2 phần Base (50 class) và Extra (10 tập nhỏ, mỗi tập chứa 5 class)

Nội dung 2: Xây dựng mô hình tiệm tiến dựa trên kiến trúc Transformer

- **Phương pháp thực hiện**
  + Chúng ta trích xuất đặc trưng của ảnh cần huấn luyện thông qua mạng pre-trained CNN
  + Đối với phần dữ liệu Base, chúng ta mong muốn huấn luyện trên mô hình Transformer để tạo ra một encoder có thể biểu diễn được đặc trưng của các lớp phần Base
  + Đối với phần Extra, mỗi Batch sẽ được đưa vào encoder để tạo ra một token mới, so sánh chúng đối với token cũ từ phần Base, từ đó so sánh được sự khác biệt giữa các đặc trưng của các lớp đối tượng mới để điều chỉnh lại encoder sao cho có thể học được các lớp mới và giữ nguyên các lớp cũ
  + Cuối cùng, sau khi cập nhật, Encoder có thể biểu diễn cả lớp đối đối tượng mới lẫn đối tượng cũ, từ đó biểu diễn thành một token mới, được đưa qua decoder để đưa ra dự đoán nhận diện vật thể



Nội dung 3: Tinh chỉnh mô hình sao cho đạt hiệu suất cao nhất

- **Phương pháp thực hiện**
  - + Đối với phần input encoder sẽ cần đánh giá trên nhiều bộ rút trích đặc trưng (Features Extractor) khác nhau từ họ CNN như VGG-16 [6], ResNet50 [7], Inception [8] để đánh giá xem bộ rút trích đặc trưng nào phù hợp với mô hình chúng ta nhất
  - + Số lượng đặc trưng được tạo ra bởi bộ encoder sẽ được tinh chỉnh để đánh giá xem một token cần biểu diễn bao nhiêu đặc trưng để có thể thực hiện việc dự đoán tốt nhất
  - + Sau khi tinh chỉnh mô hình, chúng tôi sẽ đánh giá hiệu suất của mô hình so với các phương pháp trước đó trên thang đo được sử dụng thường xuyên trong các bài báo cáo về bài toán này là mAP (Mean Average Precision) để đánh giá tính hiệu quả của mô hình

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Xây dựng một mô hình học tiên tiến nhằm đạt được kết quả vượt trội về hiệu suất hoặc thời gian so với các phương pháp hàng đầu trước đó (State of the Art). Mục tiêu của chúng tôi là đảm bảo độ chính xác của mô hình tăng lên từ 5-10% sau khi được kiểm nghiệm và đánh giá. Chúng tôi cam kết thực hiện các thử nghiệm chặt chẽ, so sánh kết quả với các phương pháp khác để chứng minh tính cạnh tranh và sự nâng cao hiệu suất của mô hình.
- Khi kết quả đạt được đủ ấn tượng, chúng tôi đặt mục tiêu đưa mô hình của mình đến mức độ có thể được chấp thuận tại các sự kiện lớn như CVPR2024. Chúng tôi tin rằng kết quả đánh giá tích cực và mức độ đóng góp của mô hình sẽ làm nổi bật nó trong cộng đồng nghiên cứu và tạo nên một bước tiến quan trọng trong lĩnh vực này.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Alex Krizhevsky: Learning Multiple Layers of Features from Tiny Images, 2009

[2]. Olga Russakovsky, *Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015

[3]. Hou, Saihui and Pan, Xinyu and Loy, Chen Change and Wang, Zilei and Lin, Dahua, "Learning a unified classifier incrementally via rebalancing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019

[4]. Yan, Shipeng, Jiangwei Xie, and Xuming He. "DER: Dynamically Expandable Representation for Class Incremental Learning.", *CPVR*, 2021. /abs/2103.16788.

[5]. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," CoRR, vol. abs/1706.03762, 2017. arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762.

[6]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Represen- tations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556 .

[7]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CoRR, vol. abs/1512.03385, 2015. arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385

[8]. C. Szegedy, W. Liu, Y. Jia, et al., "Going deeper with convolutions," CoRR, vol. abs/1409.4842, 2014. arXiv: 1409 . 4842. [Online]. Available: http://arxiv.org/abs/1409.4842

[9]. K. O'Shea and R. Nash, "An introduction to convolutional neural networks," CoRR, vol. abs/1511.08458, 2015. arXiv: 1511.08458. [Online]. Available: http://arxiv.org/abs/1511.08458 .

[10]. T. Lin, M. Maire, S. J. Belongie, et al., "Microsoft COCO: common objects in context," CoRR, vol. abs/1405.0312, 2014. arXiv: 1405.0312. [Online]. Available: http://arxiv.org/abs/1405.0312 .