THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  (ví dụ: https://www.youtube.com/watch?v=AWq7uw-36Ng)
- Link slides (dạng .pdf đặt trên Github của nhóm):
  (ví dụ: https://github.com/mynameuit/CS519.O11/TenDeTai.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: Nguyễn Đức Nhân
• MSSV: 21520373
 
	• Lớp: CS519.O11
	• Tự đánh giá (điểm tổng kết môn): 9.0/10
	• Số buổi vắng: 2
	• 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
	  ○ Viết phần toàn bộ
	  ○ Làm slide
	  ○ Làm video YouTube
	  ○ Làm poster

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

TĂNG TỐC SUY LUẬN CHO MÔ HÌNH VISION TRANSFORMER SỬ DỤNG KỸ THUẬT GỘP TOKEN VÀ HASHING

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ACCELERATING INFERENCE IN VISION TRANSFORMER MODELS USING TOKEN MERGE AND HASHING TECHNIQUES

TÓM TẮT *(Tối đa 400 từ)*

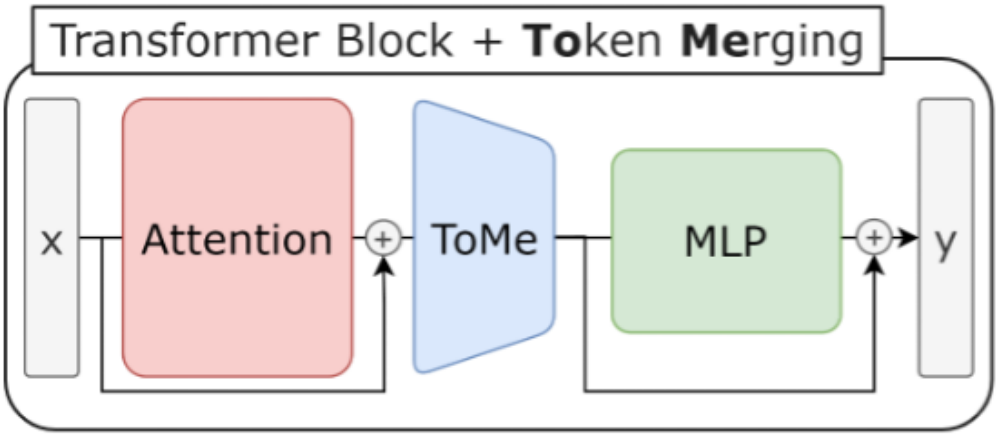
Sự ra đời của các mô hình Vision Transformer (ViT) [1] đã đánh dấu một bước tiến đáng kể về mặt hiệu suất trong các tác vụ của lĩnh vực thị giác máy tính (image recognition [1], object detection [2], image segmentation [3], …). Tuy nhiên, những yêu cầu về mặt tài nguyên phần cứng cũng như độ trễ của các mô hình Vision Transformer đặc biệt là trong quá trình suy luận đặt ra những thách thức trong việc ứng dụng vào các hệ thống thời gian thực. Nghiên cứu này tập trung vào việc đẩy nhanh tốc độ suy luận trong các mô hình Vision Transformer nhưng vẫn duy trì độ chính xác của mô hình. Phương pháp của chúng tôi tập trung vào hai kỹ thuật chính là gộp token và hàm băm với mục đích chính là giảm số lượng phép tính cần thực hiện từ đó giảm tốc độ trong quá trình suy luận. Nghiên cứu này được đánh giá trên các dataset phổ biến như ImageNet [4], Cifar100[5], Oxford-IIIT Pet[6] để chứng minh tính hiệu quả của phương pháp trên nhiều tập dữ liệu có tính chất khác nhau.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Sự ra đời của mô hình Vision Transformer [1] đã đánh dấu sự phát triển nhanh chóng của các mô hình này trong lĩnh vực thị giác máy tính. Tuy nhiên các

nghiên cứu trước đó đều bị thống trị bởi một hướng nghiên cứu chính tập trung vào việc cải tiến các module đặc trưng riêng nhằm nâng cao hiệu quả của toàn bộ kiến trúc. Có thể kể đến như Swin Transformer [7], [8] với module attention đặc trưng, LeViT [9] với module conv đặc trưng. Mục tiêu chung của các phương pháp này đều hướng tới sự hiệu quả trong kiến trúc Vision Transformer bằng các module được thiết kế riêng. Tuy nhiên điểm yếu chung của các phương pháp này là việc phải huấn luyện từ đầu và sự thiếu thân thiện trong quá trình ứng dụng, tinh chỉnh dẫn đến nhu cầu có một phương pháp tăng tốc suy luận mà không cần huấn luyện.

Một hướng tiếp cận mới triển vọng trong bài toán tăng tốc suy luận cho mô hình Vision Transformer là việc giảm thiểu số lượng token được tính toán trong quá trình suy luận từ đó làm giảm thời gian suy luận. Một số nghiên cứu tiêu biểu có thể kể đến như ToMe [11], [12] lần lượt nghiên cứu việc ứng dụng kỹ thuật gộp token cho bài toán image classification và bài toán image generation. HeatViT [13] nghiên cứu việc cắt giảm các token nhằm tối ưu tốc độ trên các thiết bị FPGA (Field-Programmable Gate Array). Các nghiên cứu trên đều cho thấy những kết quả khả quan cũng như không gian phát triển do tính mới của hướng tiếp cận.



Hình 1. Minh họa cho phương pháp ToMe [11]

Từ những lý do trên, đề tài nghiên cứu của chúng tôi tập trung vào việc kiểm thử hiệu quả của các phương pháp tăng tốc suy luận cho mô hình Vision Transformer sử dụng kỹ thuật gộp token trên nhiều dataset trong nhiều bài toán khác nhau nhằm kiểm tra tính tổng quát của phương pháp. Ngoài ra chúng tôi sẽ nghiên cứu tích hợp kỹ thuật băm cho bước gộp token nhằm tiếp tục tăng tốc cho phương pháp đã được đề xuất.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

1. Đánh giá phương pháp ToMe trên những dataset phổ biến như ImageNet, CiFar100, Oxford-IIIT Pet và các bài toán khác như object detection.
2. Triển khai kỹ thuật băm cho phương pháp ToMe và thực hiện đánh giá trên các dataset kể trên.
3. Phát triển phương pháp của chúng tôi để ứng dụng cho các loại dữ liệu khác như video, âm thanh.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Trong nghiên cứu này chúng tôi sẽ tìm hiểu các nội dung sau:

1. **Nội dung 1** - đánh giá các phương pháp đã có:
    - **Phương pháp thực hiện:** Cài đặt phương pháp ToMe cũng như các phương pháp khác cho các kiến trúc Vision Transformer như ViT, DeIT, MAE và thực hiện đánh giá trên các dataset ImageNet, Cifar100, Oxford-IIIT Pet.
    - **Kết quả dự kiến:** Biểu đồ phân tích sự đánh đổi giữa tốc độ suy luận và độ chính xác của kiến trúc trên các tập dữ liệu được sử

dụng cho đánh giá. Tài liệu mô tả, phân tích về xu hướng giảm của độ chính và so sánh với tính chất dữ liệu

2. **Nội dung 2** - tích hợp kỹ thuật băm cho phương pháp ToMe:
    - **Phương pháp thực hiện:** Tiến hành phân tích, thiết kế giải thuật và cài đặt kỹ thuật băm cho phương pháp ToMe dựa trên mã nguồn, tài liệu của nhóm tác giả. Sau đó tiến hành đánh giá trên các tập dữ liệu như ImageNet, Cifar100, Oxford-IIIT Pet.
    - **Kết quả dự kiến:** Mã nguồn và tài liệu cho phương pháp của chúng tôi. Tài liệu mô tả, phân tích về tốc độ suy luận và độ chính xác trên các tập dữ liệu được sử dụng cho đánh giá,
3. **Nội dung 3** - nghiên cứu ứng dụng với các dạng dữ liệu khác:
    - **Phương pháp thực hiện:** Tiến hành ứng dụng phương pháp của chúng tôi trong bài toán video retrieval.
    - **Kết quả dự kiến:** Bảng kết quả mô tả độ chính xác cũng như tốc độ suy luận được đánh giá trên các dataset phổ biến cho Video retrieval như tập dữ liệu AI Challenge 2023.

KẾT QUẢ MONG ĐỢI

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

1. Mã nguồn, tài liệu kỹ thuật cho việc cài đặt các phương pháp tăng tốc suy luận cho mô hình Vision Transformer. Mã nguồn cài đặt cho việc đánh giá các phương pháp tăng tốc suy luận trên các tập dataset đã liệt kê.
2. Bảng kết quả, thông tin kỹ thuật cho các thực nghiệm được tiến hành ở nội dung 1, 2, 3.
3. So sánh kết quả giữa các phương pháp được thực nghiệm và ứng dụng

truy vấn dữ liệu có ứng dụng phương pháp trong nghiên cứu.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby:
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

[2]. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko:
End-to-End Object Detection with Transformers. ECCV (1) 2020: 213-229

[3]. Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, Cordelia Schmid:
Segmenter: Transformer for Semantic Segmentation. CoRR abs/2105.05633 (2021)

[4]. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei:
ImageNet: A large-scale hierarchical image database. CVPR 2009: 248-255

[5]. Alex Krizhevsky:
Learning Multiple Layers of Features from Tiny Images.

[6]. Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, C. V. Jawahar:
Cats and dogs. CVPR 2012: 3498-3505

[7]. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo:
Swin transformer: Hierarchical vision transformer using shifted windows. ICCV 2021.

[8]. Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng:
Swin transformer v2: Scaling up capacity and resolution. CVPR 2022a.

[9]. Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Herve J ´ egou, ´ and Matthijs Douze:
Levit: a vision transformer in convnet’s clothing for faster inference. ICCV

2021.

[10]. Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou:

Training data-efficient image transformers & distillation through attention. ICML 2021: 10347-10357

[11]. Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, Judy Hoffman:

Token Merging: Your ViT But Faster. ICLR 2023

[12]. Daniel Bolya, Judy Hoffman:

Token Merging for Fast Stable Diffusion. CVPR Workshops 2023: 4599-4603

[13]. Peiyan Dong, Mengshu Sun, Alec Lu, Yanyue Xie, Kenneth Liu, Zhenglun Kong, Xin Meng, Zhengang Li, Xue Lin, Zhenman Fang, Yanzhi Wang:

HeatViT: Hardware-Efficient Adaptive Token Pruning for Vision Transformers. HPCA 2023: 442-455