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

Nguyễn Đức Nhân - 21520373

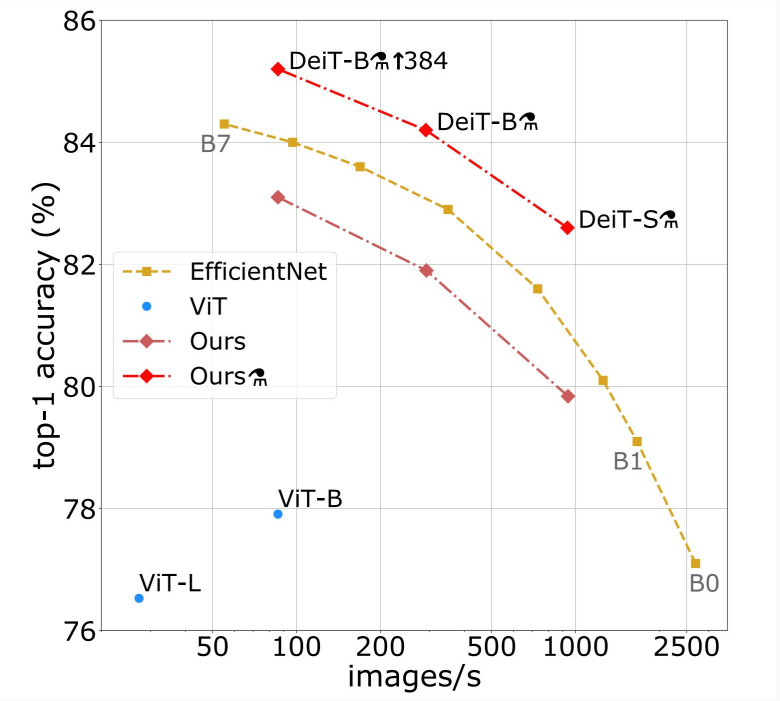
Tóm tắt

- Lớp: CS519.O11
- Link Github của nhóm:
- Link YouTube video:

Nguyễn Đức Nhân

Giới thiệu

- Các mô hình Vision Transformer đã đạt được những kết quả ấn tượng vượt trội so với các mô hình dựa trên kiến trúc CNN.
- Các mô hình Vision Transformer bộc lộ những nhược điểm rõ thấy như tốc độ suy luận thấp, cần nhiều tài nguyên phần cứng
- Nhu cầu về một phương pháp giúp tăng tốc suy luận nhưng vẫn giữ độ chính xác ở mức chấp nhận được.



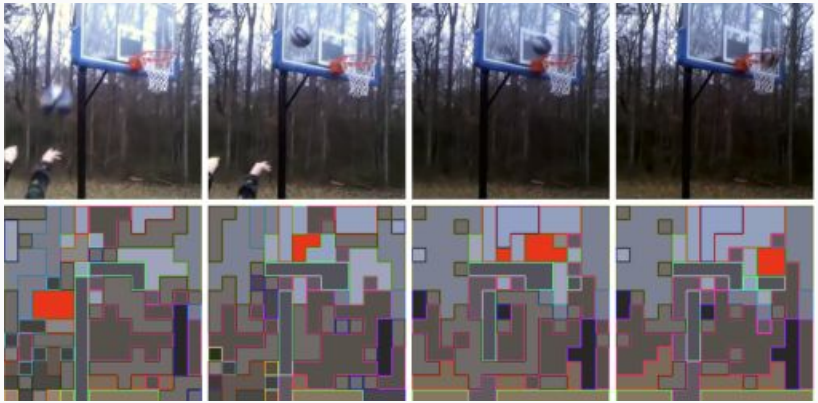
Hình 1. Biểu đồ giữa độ chính xác và tốc độ suy luận của mô hình ViT, EfficientNet, DeIT [1]

Mục tiêu

1. Nghiên cứu, thực nghiệm các phương pháp tăng tốc suy luận dựa trên kỹ thuật gộp token như ToMe [2] cho bài toán image classification, object detection.
2. Triển khai các kỹ thuật giảm thiểu tốc độ như kỹ thuật hashing và tích hợp chúng vào phương pháp ToMe.
3. Mở rộng phạm vi thực nghiệm cho các phương pháp tăng tốc suy luận dựa trên kỹ thuật gộp token cho các loại dữ liệu khác như video, âm thanh.

Nội dung và Phương pháp

1. Đánh giá các phương pháp đã có và kết hợp kỹ thuật hashing cho bài toán image classification trên các dataset như ImageNet [3], Cifar100[4], Oxford-IIIT Pet[5]
2. Nghiên cứu ứng dụng các phương pháp đã được nghiên cứu ở nội dung trên cho dữ liệu dạng video, âm thanh.



Hình 2. Minh họa các token sau khi áp dụng kỹ thuật gộp token cho bài toán video retrieval

Kết quả dự kiến

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 ImageNet, Cifar100, Oxford-IIIT Pet.
2. Bảng kết quả, thông tin kỹ thuật cho các thực nghiệm áp dụng kỹ thuật gộp token và hashing cho bài toán image classification và bài toán video retrieval.
3. Paper tại hội nghị ICCV.

Tài liệu tham khảo

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

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

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

Token Merging: Your ViT But Faster. ICLR 2023

[3]. 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

[4]. Alex Krizhevsky:

Learning Multiple Layers of Features from Tiny Images.

[5]. Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, C. V. Jawahar:

Cats and dogs. CVPR 2012: 3498-3505