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

Nguyễn Đức Nhân [1,2]

[1]21520373@gm.uit.edu.vn [2]Trường Đại học Công nghệ Thông tin - ĐHQG TP.HCM

Vấn đề chính

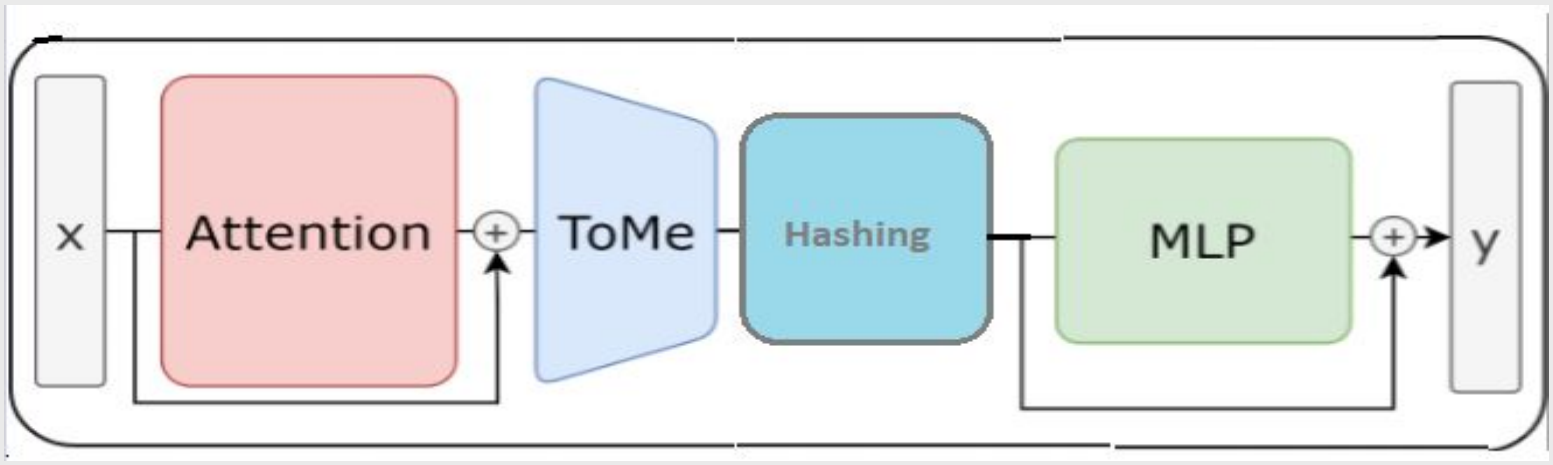
Chúng tôi đề xuất một phương pháp tăng tốc suy luận trên mô hình Vision Transformer:

- Tăng tốc độ suy luận lên đến 1.5 lần với độ chính xác chỉ giảm 1%.
- Phương pháp của chúng tôi không cần huấn luyện thêm và mọi tăng tốc đều diễn ra tự động.
- Được đánh giá trên những bộ dữ liệu phổ biến như ImageNet, Cifar100, Oxford-IIIT Pet

Lý do chọn đề tài

- Các mô hình Vision Transformer đã đạt được những kết quả rất khả quan trong nhiều tác vụ thị giác máy tính nhưng gặp khó khăn vì yêu cầu về tài nguyên tính toán cũng như tốc độ suy luận chậm.
- Các nghiên cứu trước thường tập trung vào việc áp dụng các module đặc trưng nhưng đánh đổi là khó ứng dụng.
- Nhu cầu về một phương pháp có sự trao đổi về độ chính xác thấp cũng như không cần huấn luyện lại mô hình.

Tổng quan về phương pháp gộp token kết hợp với hashing



Mô tả

1. Nội dung:

- Nghiên cứu và thực nghiệm các phương pháp tăng tốc mô hình Vision Transformer bằng kỹ thuật gộp token và hashing.
- Chứng minh tính hiệu quả của phương pháp đối với các mô hình Vision Transformer.
- Ứng dụng phương pháp được nghiên cứu cho các kiểu dữ liệu khác như video, âm thanh.

2. Phương pháp:

- Tiến hành nghiên cứu, phân tích, thiết kế giải thuật để tích hợp kỹ thuật hashing và phương pháp gộp token dựa trên các nghiên cứu trước đó.
- Thực hiện các đánh giá cho các phương pháp đã có cũng như phương pháp được nghiên cứu trên các tập dữ liệu phổ biến như ImageNet, Cifar100, …
- Ứng dụng kỹ thuật tăng tốc suy luận cho mô hình Vision Transformer trên dữ liệu là video.

3. Kết quả mong đợi:

- 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.
- 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.

NII

Nguyễn Đức Nhân – Trường Đại học Công nghệ Thông tin - ĐHQG TPHCM

TEL : 0367737219 Email : 21520373@gm.uit.edu.vn