MÔ HÌNH TRANSFORMER CHO NHẬN DẠNG HÌNH ẢNH TRÊN QUY MÔ LỚN

Phạm Nguyễn Xuân Trường[1,2]

Lê Đăng Khoa [1,3]

[1] Trường Đại học Công nghệ Thông tin ĐHQG TP.HCM

[2] 20520835@gm.uit.edu.vn

[3] 21522222@gm.uit.edu.vn

What ?

- Tìm hiểu về **Transformer** và áp dụng vào bài toán **Image recognition.**
- Pre-train mô hình **Vision Transformer** với tập dữ liệu lớn (**JFT**, **ImageNet**) và so sánh với **ResNet**, **EfficientNet** để đánh giá hiệu suất và chất lượng.
- Xây dựng chương trình ứng dụng minh họa sử dụng **ViT.**

Why ?

- Nhận dạng hình ảnh là nhiệm vụ quan trọng trong thị giác máy tính.
- CNNs là mô hình học sâu phổ biến nhất cho nhận dạng hình ảnh, nhưng có giới hạn trong việc bắt được phụ thuộc xa và thời gian đào tạo dài, chi phí tính toán cao khi tập dữ liệu và kiến trúc mô hình tăng lên.
- Phương pháp mới **Vision Transformers (ViT)** được đề xuất để hiệu quả hơn ở các vấn đề của CNNs trong việc nhận dạng hình ảnh ở quy mô lớn hơn.

Overview

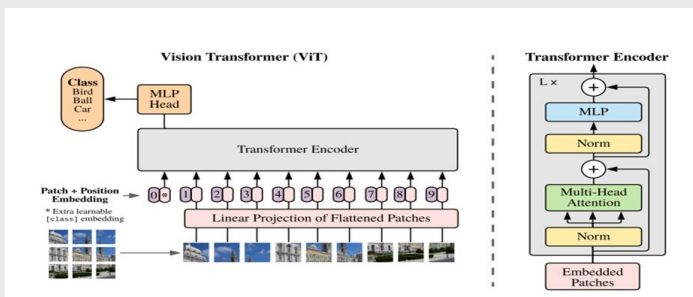
Description

1. Nội dung

- Tìm hiểu về kiến trúc của **Transformer.**
- Tìm hiểu về bài toán **Image recognition.**
- Tìm hiểu về các mô hình Transformer được áp dụng cho Image recognition bằng cách trả lời câu hỏi: "**Biến đổi ảnh đầu vào như thế nào để áp dụng được với mô hình Transformer?**". Nghiên cứu và phân tích các bài toán sử dụng tích hợp transformer để trả lời câu hỏi (Wang et al., 2018; Carion et al., 2020, Ramachandran et al., 2019)
- Nghiên cứu và tìm hiểu tập dữ liệu lớn **ImageNet** và **JFT-300M**, sử dụng **ResNet** và **EfficientNet** để đánh giá hiệu suất.

2. Phương pháp:

- Áp dụng encoder của Transformer trực tiếp lên hình ảnh bằng cách chia hình thành các ảnh nhỏ và sử dụng embedding tuyến tính của chúng.

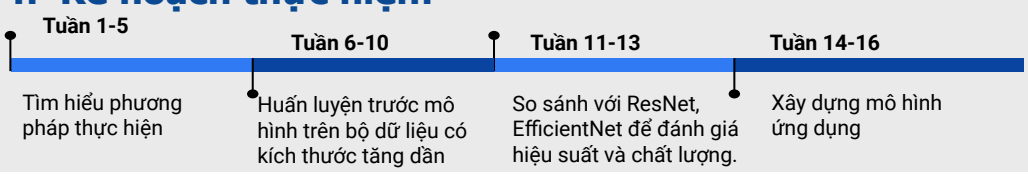


- Chúng tôi cũng thêm một embedding học được vào đầu chuỗi các patch và positions embedding để giữ lại thông tin vị trí.
- Các mô hình ViT sẽ được pre-train trên các tập dữ liệu có kích thước tăng dần: ImageNet, ImageNet-21k và JFT300M.
- Chúng tôi sẽ kiểm tra khả năng mở rộng của mô hình (**scalability**) trên tập dữ liệu **ILSVRC-2012 ImageNet** với 1.000 lớp và so sánh kết quả với mô hình **Big Transfer** và **Noisy Student**. BiT sử dụng học chuyển giao có giám sát với ResNets, trong khi Noisy Student sử dụng EfficientNet được huấn luyện bằng phương pháp học bán giám sát trên ImageNet và JFT300M mà không có nhãn.
- Sau khi pre-train ViT, chúng tôi sẽ dự định sử dụng kết quả và tập data **ILSVRC-2012 ImageNet** để trực quan hóa dữ liệu. Chương trình có thể được đăng trên Internet.

3. Kết quả dự kiến:

- Nếu kết quả đạt được thành công - nhận dạng hình ảnh ở quy mô lớn nhanh chóng và chính xác hơn so với CNN - thì chúng tôi dự định đặt tên cho phương pháp này là ViT (Vision Transformer).
- Chúng tôi dự kiến công bố:
  - 01 bài báo được đăng trên Hội nghị thường niên về Xử lý Ngôn ngữ Tự nhiên và Công nghệ Thông tin (ACL-IJCNLP).
  - 01 mô hình pre-trained trên Google Research

4. Kế hoạch thực hiện:



Nhóm:
Github: CS519.N11 (Github)

Phạm Nguyễn Xuân Trường – UIT
TEL:

Lê Đăng Khoa – UIT
TELL: