BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS519 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS519.N11 - CS519.N11.KHCL

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

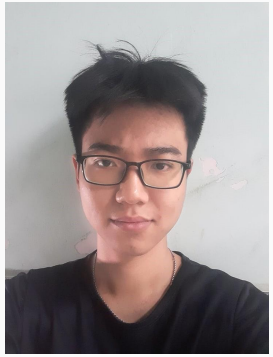
MÔ HÌNH TRANSFORMER CHO NHẬN DẠNG HÌNH ẢNH TRÊN QUY MÔ LỚN

Phạm Nguyễn Xuân Trường – 20520835

Lê Đăng Khoa - 21522222

Tóm tắt

- Lớp: CS519.N11
- Link Github của nhóm: https://github.com/pnxuantruong/CS519.N11
- Link YouTube video:



Phạm Nguyễn Xuân Trường - 20520835

Lê Đăng Khoa - 21522222

Giới thiệu

Khi tập dữ liệu hình ảnh và kiến trúc mô hình tăng lên, đối với:

- Mạng nơ-ron tích chập (CNNs)
  - Thời gian đào tạo dài
  - Chi phí tính toán cao
  - Có giới hạn trong khả năng bắt được phụ thuộc xa trong hình ảnh, dẫn đến độ chính xác giảm trên các nhiệm vụ nhận dạng hình ảnh phức tạp
- Mô hình **Vision Transformers**
  - Sử dụng các cơ chế **self-attention** để xử lý dữ liệu tuần tự
  - Bắt được phụ thuộc xa trong hình ảnh một cách hiệu quả hơn và với tính toán song song
  - Nhận dạng hình ảnh nhanh và chính xác hơn ở quy mô lớn hơn

Giới thiệu

- **Input:** Một hình ảnh
- **Ouput:** Các đặc trưng cấp cao (high-level features)

Mục tiêu

- Tìm hiểu về Transformer và áp dụng vào bài toán Image recognition.
- Tiền huấn luyện (Pre-train) mô hình Vision Transformer với tập dữ liệu lớn (JFT, ImageNet) và so sánh với ResNet, EfficientNet [5] [6] để đánh giá hiệu suất và chất lượng.
- Xây dựng chương trình ứng dụng minh họa sử dụng ViT.

Nội dung

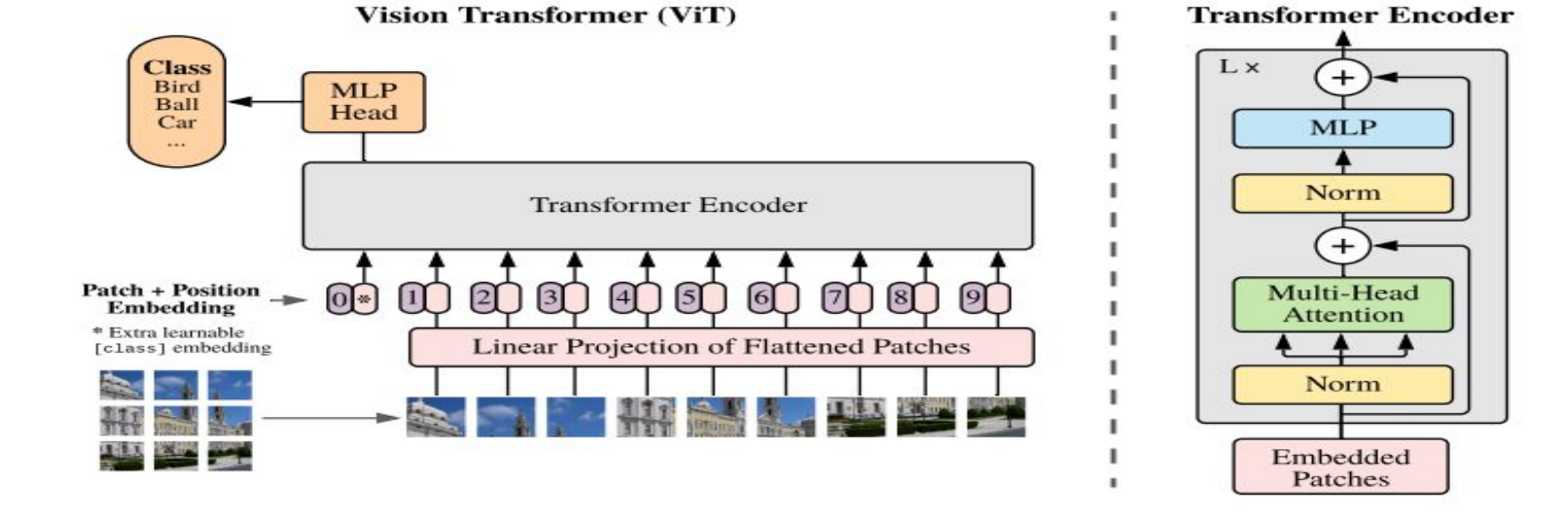
- Tìm hiểu về kiến trúc của Transformer.
- Tìm hiểu về bài toán Image recognition.
- Tìm hiểu về các mô hình Transformer được áp dụng cho Image recognition bằng cách trả lời câu hỏi: “**Biến đổi ảnh đầu vào như thế nào để áp dụng được với mô hình Transformer?**”. Nghiên cứu và phân tích các bài toán sử dụng tích hợp transformer để trả lời câu hỏi (Wang et al., 2018 [2]; Carion et al., 2020 [3], Ramachandran et al., 2019 [4])
- Nghiên cứu và tìm hiểu tập dữ liệu lớn ImageNet và JFT-300M, sử dụng ResNet [5] và EfficientNet [6] để đánh giá hiệu suất.

Phương pháp

- Transformer là một kiến trúc mạng nơ-ron thần kinh sử dụng cơ chế attention để xử lý các chuỗi dữ liệu như văn bản, âm thanh và hình ảnh được giới thiệu trong [1]
- Bài toán Image recognition là một bài toán trong lĩnh vực Computer Vision, có nhiệm vụ phân loại các hình ảnh vào các lớp khác nhau.
- hình ảnh được chúng tôi áp dụng encoder của Transformer trực tiếp lên bằng cách chia hình ảnh thành các ảnh nhỏ và sử dụng embedding tuyến tính của chúng như là đầu vào cho Transformer, giống như cách xử lý token trong ứng dụng NLP.

Phương pháp

- Giống như [class] token của BERT [7], chúng tôi thêm một embedding có thể học vào đầu chuỗi các patch đã được nhúng để tạo ra biểu diễn hình ảnh (1 class). Chúng tôi cũng thêm embedding vị trí vào các embedding của patch để giữ lại thông tin vị trí.



Phương pháp

- Các mô hình ViT được chúng tôi huấn luyện trước (pre-train) trên các tập dữ liệu có kích thước tăng dần: ImageNet, ImageNet-21k và JFT300M. Để kiểm tra khả năng mở rộng của mô hình (**scalability**), chúng tôi dự kiến sử dụng tập dữ liệu **ILSVRC-2012 ImageNet** với 1.000 lớp và 1,3 triệu hình ảnh là tập con của ImageNet-21k [8] và JFT [9].
- Chúng tôi dự định sẽ so sánh kết quả với 2 mô hình tiên tiến (state-of-the-art): Big Transfer (BiT) [10] và Noisy Student [6]. 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.

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

Tài liệu tham khảo

- [1]. D. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In NIPS, pages 5998–6008, 2017.
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.
- [3] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In CVPR, 2018.
- [4] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In NeurIPS, 2019.
- [5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778).
- [6] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In CVPR, 2020