


# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://youtu.be/St8n8ktq1kk>
- Link slides (dạng .pdf đặt trên Github của nhóm):  
<https://github.com/pnxuantruong/CS519.N11/blob/main/DoAnCuoiKy/CS519.N11.Slide.FinalReport.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

<ul style="list-style-type: none"><li>● Họ và Tên: Phạm Nguyễn Xuân Trường</li><li>● MSSV: 20520835</li></ul> 	<ul style="list-style-type: none"><li>● Lớp: CS519.N11</li><li>● Tự đánh giá (điểm tổng kết môn): 8/10</li><li>● Số buổi vắng: 2</li><li>● Số câu hỏi QT cá nhân: 8</li><li>● Số câu hỏi QT của cả nhóm: 3</li><li>● Link Github: <a href="https://github.com/pnxuantruong/CS519.N11">https://github.com/pnxuantruong/CS519.N11</a></li><li>● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none"><li>○ Lên ý tưởng</li><li>○ Viết phần Tóm tắt, Giới thiệu, Mục tiêu</li><li>○ Thuyết trình, chỉnh sửa video</li></ul></li></ul>
<ul style="list-style-type: none"><li>● Họ và Tên: Lê Đăng Khoa</li><li>● MSSV: 21522222</li></ul>	<ul style="list-style-type: none"><li>● Lớp: CS519.N11</li><li>● Tự đánh giá (điểm tổng kết môn): 8/10</li><li>● Số buổi vắng: 2</li><li>● Số câu hỏi QT cá nhân: 8</li><li>● Số câu hỏi QT của cả nhóm: 3</li></ul>



- Link Github:  
<https://github.com/pnxuantruong/CS519.N11>
- Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
  - Lên ý tưởng
  - Viết phần Nội dung và phương pháp
  - Làm slide, poster, chỉnh sửa video

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

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

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

## TÓM TẮT (*Tối đa 400 từ*)

Trong lĩnh vực xử lý ngôn ngữ tự nhiên, kiến trúc **Transformer** [1] đã trở thành tiêu chuẩn thực tế. Tuy nhiên, trong thị giác máy tính, việc áp dụng **attention** của Transformer vẫn còn hạn chế. Các phương pháp hiện tại thường kết hợp attention với mạng tích chập hoặc thay thế một số thành phần của mạng tích chập (CNN) với attention, nhưng vẫn giữ nguyên cấu trúc tổng thể của mạng tích chập. Tuy nhiên, chúng tôi tin rằng việc phụ thuộc vào mạng tích chập không cần thiết và việc áp dụng transformer trực tiếp lên các chuỗi ảnh có thể mang lại kết quả tốt trên các tác vụ phân loại hình ảnh. Vì vậy chúng tôi đề xuất phương pháp mới là **Vision Transformer** (ViT) sử dụng cơ chế **self-attention** để trích xuất đặc trưng từ hình ảnh. Chúng tôi dự định tiền huấn luyện mô hình Vision Transformer (ViT) trên lượng lớn dữ liệu và chuyển giao mô hình này sang nhiều bộ kiểm tra nhận dạng hình ảnh trung bình hoặc nhỏ (bao gồm ImageNet, CIFAR-100, VTAB, vv). Kết quả đạt được sẽ được so sánh với các mạng tích chập hiện đại để kiểm tra hiệu suất và tài nguyên tính toán khi huấn luyện.

## GIỚI THIỆU (*Tối đa 1 trang A4*)

Nhận dạng hình ảnh là một nhiệm vụ quan trọng trong thị giác máy tính, và học sâu đã cho thấy sự thành công đáng kể trong việc đạt được độ chính xác cao trong các nhiệm vụ nhận dạng hình ảnh. Mạng nơ-ron tích chập (CNNs) là mô hình học sâu được sử dụng rộng rãi nhất cho các tác vụ nhận dạng hình ảnh. Tuy nhiên, khi tập dữ liệu hình ảnh và kiến trúc mô hình tăng lên, độ phức tạp tính toán của CNN cũng tăng,

dẫn đến thời gian đào tạo dài và chi phí tính toán cao. Ngoài ra, CNN có giới hạn trong khả năng bắt được phụ thuộc xa trong hình ảnh, có thể dẫn đến độ chính xác giảm trên các nhiệm vụ nhận dạng hình ảnh phức tạp.

Trong khi đó, các transformer ban đầu được phát triển cho các nhiệm vụ xử lý ngôn ngữ tự nhiên (NLP) nhưng gần đây đã cho thấy tiềm năng lớn trong việc nhận dạng hình ảnh ở quy mô lớn [2][3][4]. Chúng tôi đề xuất phương pháp mới có tên **Vision Transformers** (ViT) là kiến trúc mạng nơ-ron sử dụng các cơ chế **self-attention** để xử lý dữ liệu tuần tự, cho phép có thể bắt được phụ thuộc xa trong hình ảnh một cách hiệu quả hơn và với tính toán song song, có thể dẫn đến việc nhận dạng hình ảnh nhanh hơn và chính xác hơn ở quy mô lớn hơn.

Đề xuất của chúng tôi tạo cơ hội để cải thiện hiệu quả và độ chính xác của các mô hình học sâu cho nhiệm vụ quan trọng này. Với đề xuất này, chúng tôi sẽ khám phá việc sử dụng transformers cho việc nhận dạng hình ảnh ở quy mô lớn và điều tra tiềm năng ưu điểm của chúng so với CNNs truyền thống. Đề xuất được hy vọng mang lại tác động tiềm năng của đối với lĩnh vực thị giác máy tính.

**Input:** Một hình ảnh

**Output:** Các đặc trưng cấp cao (high-level features)



(Hình ảnh minh họa khi kết hợp đặc trưng cao cấp để nhận diện ảnh)

## MỤC TIÊU

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

1. Tìm hiểu về Transformer và áp dụng vào bài toán Image recognition.
2. 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.
3. Xây dựng chương trình ứng dụng minh họa sử dụng ViT.

## NỘI DUNG VÀ PHƯƠNG PHÁP

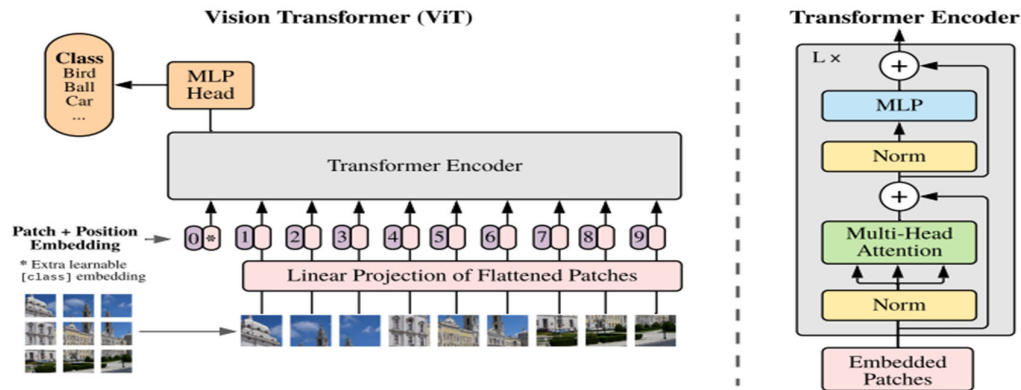
*(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)*

### 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 [2]; Carion et al., 2020 [3], Ramachandran et al., 2019 [4])
- Nghiên cứu và tìm hiểu tập dữ liệu lớn ImageNet [8] và JFT [9], sử dụng ResNet [5] và EfficientNet [6] để đánh giá hiệu suất.

### 2. 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.
- 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.



### *Mô hình dự kiến của chúng tôi*

- Các mô hình ViT được 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 MONG ĐỢI**

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

1. 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).
2. 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
3. Một chương trình trực quan của ViT trên github, kaggle, ...

### **TÀI LIỆU THAM KHẢO** (*Định dạng DBLP*)

- [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.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019.
- [8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.
- [9] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In ICCV, 2017.
- [10] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. In ECCV, 2020.