**Tóm tắt**

* Vision Transformer (ViT) là một kiến ​​trúc mạng nơ-ron sử dụng cấu trúc Transformer để xử lý hình ảnh. Thay vì sử dụng mạng nơ-ron tích chập như CNN, ViT chia ảnh thành các miền nhỏ gọi là patch và biến đổi chúng thành các vectơ, sau đó đưa vào một mạng Transformer để xử lý. ViT đã chứng minh khả năng cạnh tranh với CNN trên nhiều tác vụ thị giác máy tính, như nhận dạng hình ảnh và phân loại. Mặc dù đòi hỏi nhiều tài nguyên tính toán hơn, ViT mang lại một cách tiếp cận linh hoạt hơn cho xử lý hình ảnh, cho phép áp dụng các khái niệm từ Transformer, đã chứng tỏ hiệu suất tốt trên nhiều tập dữ liệu và tác vụ. [1]

**Giới thiệu mô hình**

* ViT đang có những bước tiến quan trọng trong ứng dụng thực tế qua các lĩnh vực khác nhau. Trong lĩnh vực y tế, ViT được sử dụng cho các nhiệm vụ phân tích hình ảnh y khoa như nhận diện các bất thường trong các bức ảnh X-quang và MRI, thể hiện tiềm năng hỗ trợ các chuyên gia y tế trong việc chẩn đoán và lập kế hoạch điều trị. Trong các hệ thống lái tự động, ViT được sử dụng cho việc hiểu cảnh và nhận diện đối tượng, đóng góp vào việc tạo ra một giao thông an toàn và hiệu quả hơn. Hơn nữa, trong lĩnh vực nông nghiệp, ViT hỗ trợ trong việc giám sát mùa vụ và phát hiện bệnh qua phân tích hình ảnh từ không gian, giúp cho các nông dân có thể ra quyết định dựa trên dữ liệu để tăng sản lượng và chất lượng của cây trồng. Tính linh hoạt và hiệu suất ổn định của nó khiến cho ViT trở thành một giải pháp hứa hẹn để giải quyết các nhiệm vụ thị giác phức tạp trong các bối cảnh thực tế, mở đường cho sự tiến bộ trong nhiều ngành công nghiệp khác nhau. [1]
* Mô hình Transformer, một mô hình mang tính cách mạng, được Google giới thiệu năm 2017 trong một bài nghiên cứu “Attention is all you need”. Transformer chủ yếu được tạo ra cho các nhiệm vụ NLP. Khác với RNN những điểm sau đây:
  + RNNs xử lý tuần tự. Nếu chúng ta có một input là một câu thì RNNs chỉ xử lý một chữ tại một thời điểm. Còn Transformer là phi tuần tự, sẽ xử lý tất cả các từ cùng một lúc.
  + Điều làm cho Transformer trở nên đặc biệt và vượt trội hơn RNNs là sử dụng cơ chế self-attetion. Cơ chế này cho phép Transformer có thể hiểu được ngữ nghĩa của cả câu và tìm ra sự liên kết giữa các từ trong câu nhưng RNNs thì không.
  + Positional Embeddings: Transformer sử dụng để lưu trữ thông tin về vị trí của các từ trong câu
* Từ những sự khác biệt đã được nêu ra ở phía trên, Transformer đã trở nên phổ biến ngay lập tức và đã vượt qua RNNs trong các nhiệm vụ NLP. [2]
* Trong khi mô hình Transformer trở nên cực kỳ phổ biến trong các nhiệm vụ NLP. Một bài báo có tên “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” [3] ứng dụng của Transformer trong computer vision vẫn còn giới hạn. Trong computer vision, attetion được áp dụng cùng với mạng tích chập hoặc được sử dụng để thay thế một số thành phần nhất định của mạng tích chập trong khi vẫn giữ nguyên cấu trúc tổng thể của chúng. Chúng tôi cho thấy rằng sự phụ thuộc vào CNN này là không cần thiết và chỉ sử dụng Transformer được áp dụng trực tiếp vào hình ảnh có thể thực hiện rất tốt các nhiệm vụ phân loại hình ảnh. Khi được đào tạo trước về lượng lớn dữ liệu và được chuyển sang nhận dạng hình ảnh cỡ vừa hoặc nhỏ (ImageNet, CIFAR-100, VTAB, v.v.), Vision Transformer (ViT) đạt được kết quả xuất sắc so với mạng tích chập trong khi yêu cầu ít tài nguyên tính toán hơn để đào tạo.[3]

**[1]:**[**https://chat.openai.com/share/0fa41c01-8219-4676-a11a-05476f13d2ec**](https://chat.openai.com/share/0fa41c01-8219-4676-a11a-05476f13d2ec)

**[2]:**[**https://sanchman21.medium.com/evolution-of-transformers-part-1-faac3f19d780**](https://sanchman21.medium.com/evolution-of-transformers-part-1-faac3f19d780)

**[3]:**<https://doi.org/10.48550/arXiv.2010.11929>