Report paper: AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Giáp Ngọc Hiệu

**Tóm tắt nội dung**

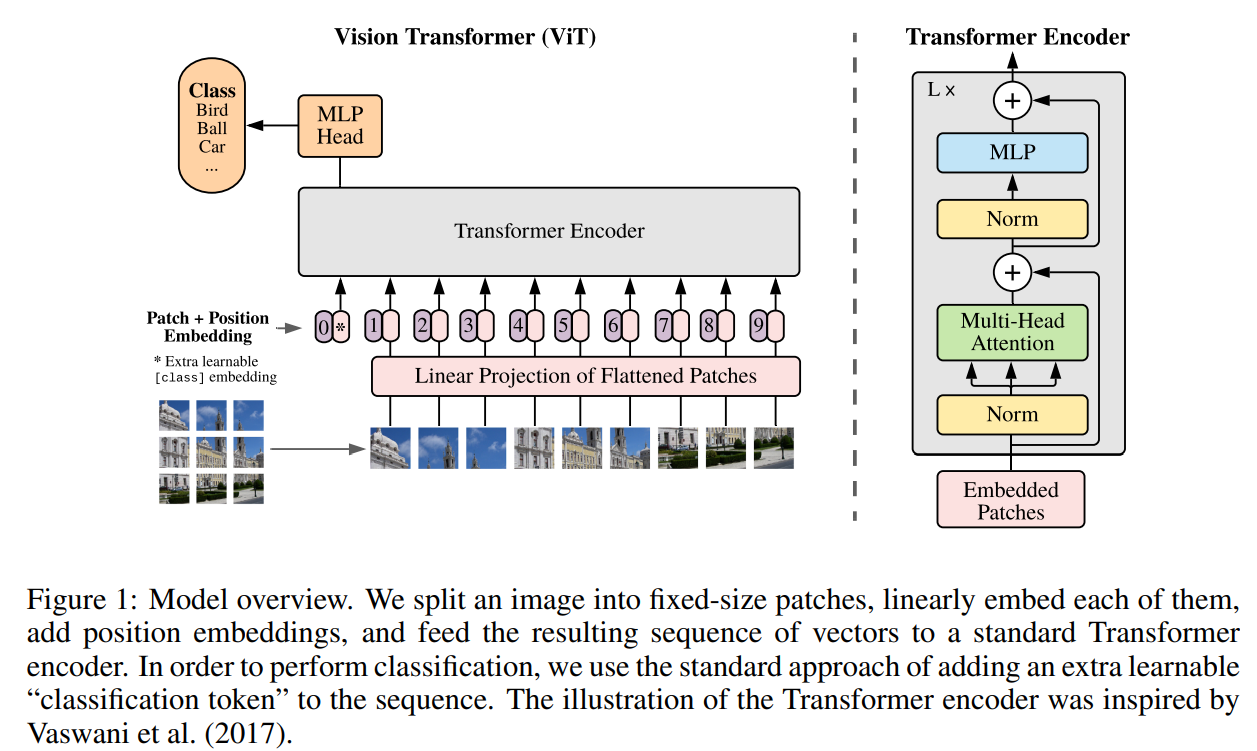
Bài report này là phần tóm tắt lại ý tưởng, nội dung chính, những điều làm được của paper

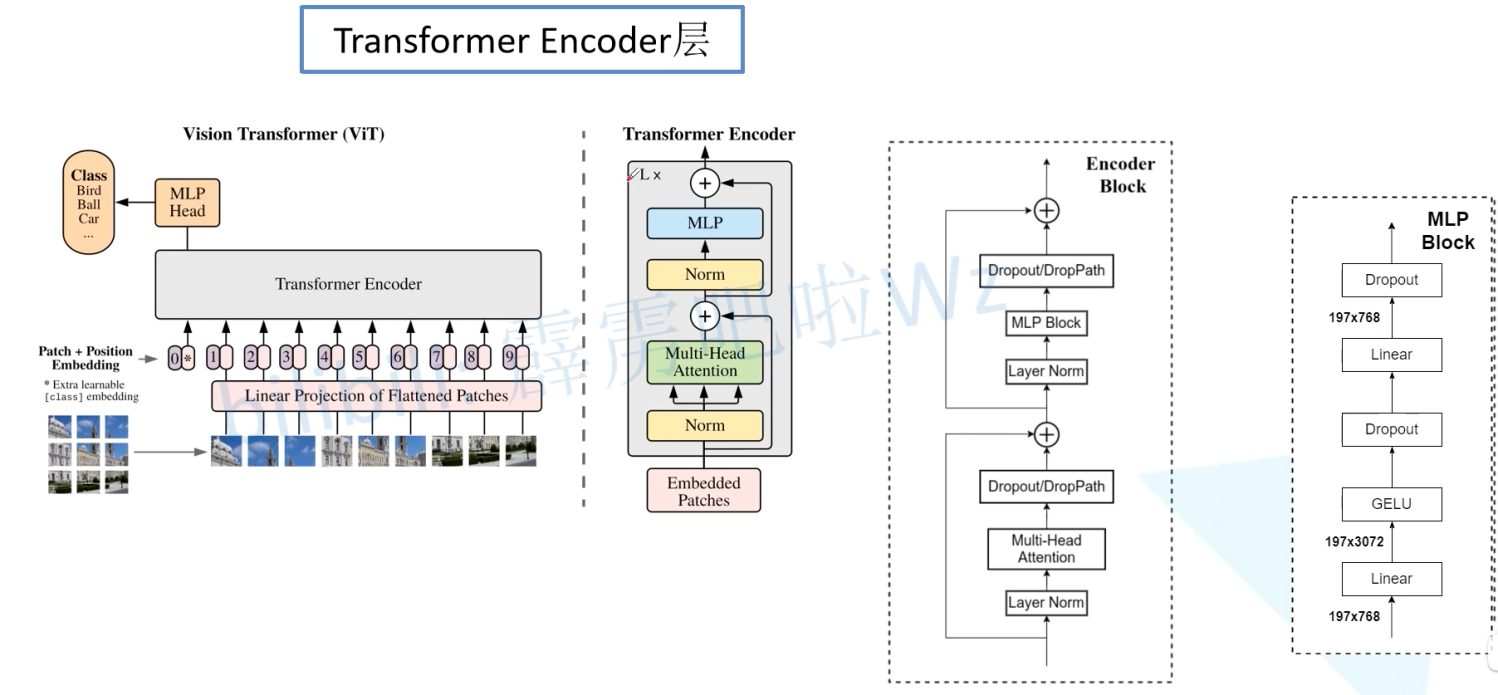
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

1. **Ý tưởng**

Mặc dù kiến ​​trúc Transformer đã trở thành tiêu chuẩn cho các tác vụ xử lý ngôn ngữ tự nhiên nhưng các ứng dụng của nó vào thị giác máy tính vẫn còn hạn chế. Trong thị giác máy tính, attention được áp dụng cùng với convolutional networks hoặc được sử dụng để thay thế một số thành phần nhất định của convolutional networks trong khi vẫn giữ nguyên cấu trúc tổng thể của chúng. Tác giả cho thấy rằng sự phụ thuộc vào CNN này là không cần thiết và một transformer thuần túy được áp dụng trực tiếp vào chuỗi các bản vá hình ảnh có thể thực hiện rất tốt các nhiệm vụ phân loại hình ảnh.

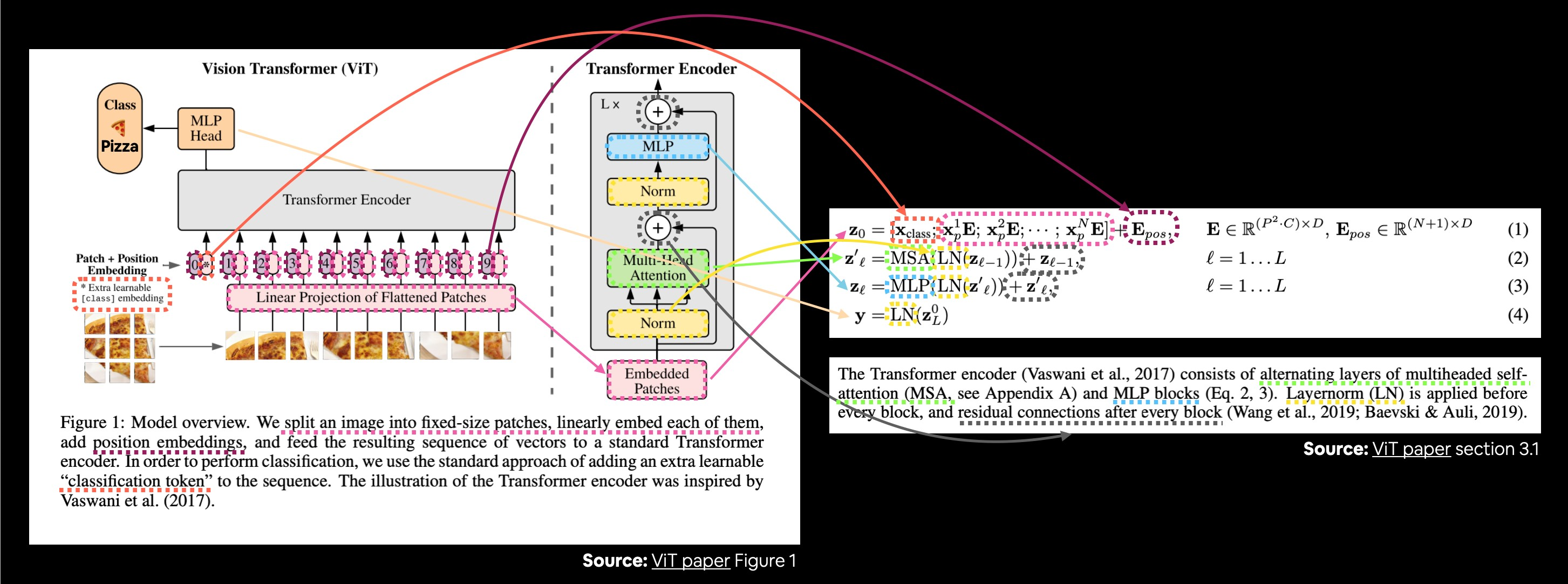
1. **Phương pháp**

****

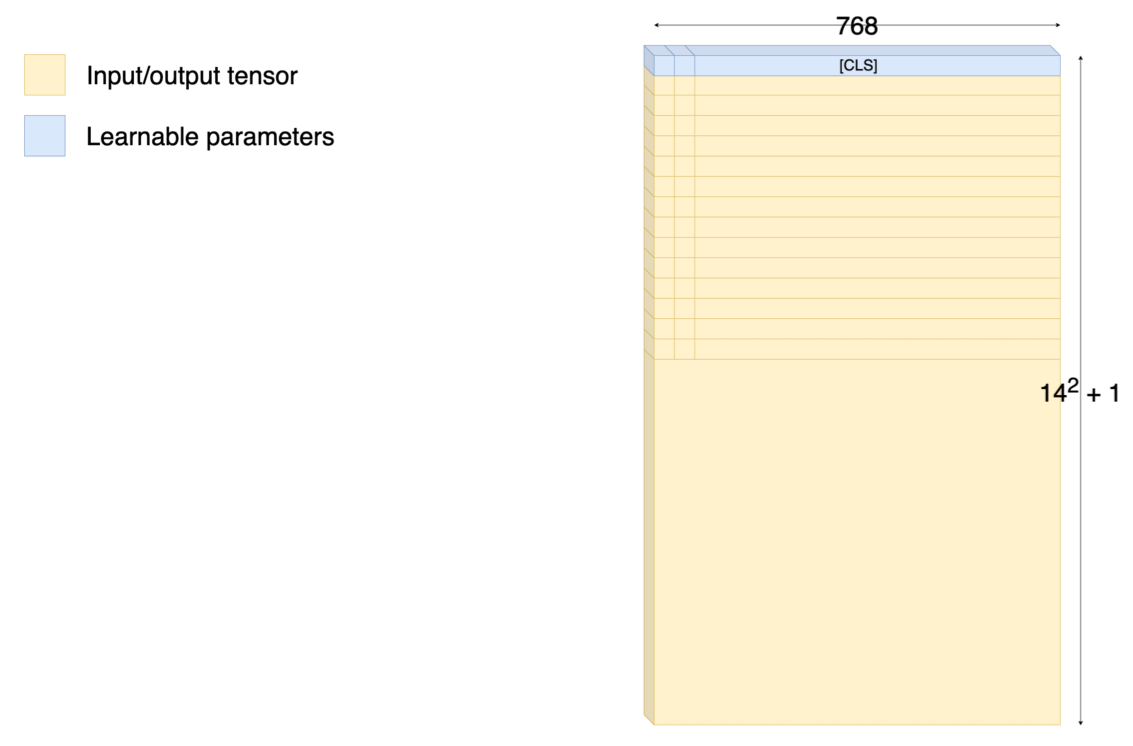


* 1. **Vision Transformer (ViT)**
* Để xử lý hình ảnh 2D, hình ảnh được reshape lại từ H×W×C thành một chuỗi các patch 2D phẳng xp, có hình dạng N×(P²×C), trong đó (H, W) là độ phân giải của hình ảnh gốc , C là số lượng kênh, (P, P) là độ phân giải của từng patch hình ảnh và N=HW/P² là số lượng patch thu được.

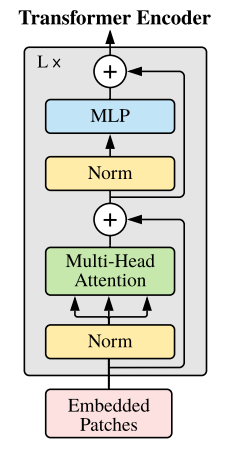




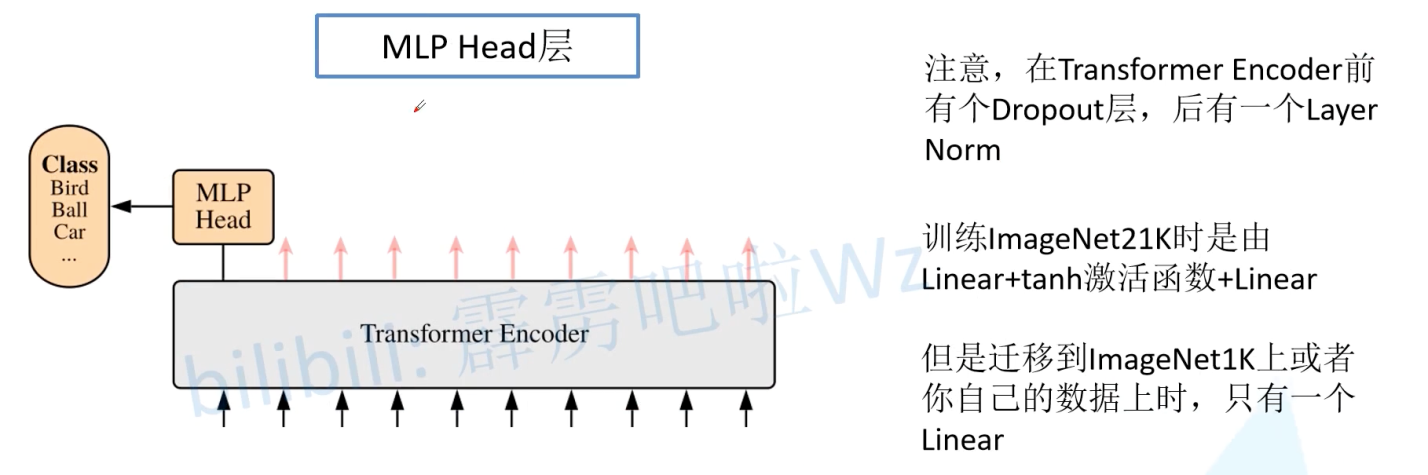
* (1): Transformer sử dụng latent vector có kích thước D không đổi thông qua tất cả các lớp của nó, do đó, các patch được làm phẳng và ánh xạ tới các kích thước D bằng phép chiếu tuyến tính có thể huấn luyện được. Đầu ra của phép chiếu này dưới dạng patch embeddings
* Tương tự như [class] token ([CLS] token) của BERT, một learnable embedding được được thêm vào trước chuỗi các embedded patches



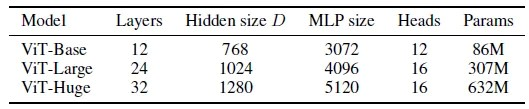
* Classification head chỉ sử dụng encoding [CLS] token từ encoding layer cuối cùng. Bằng cách này, phần tử này sẽ buộc phải học cách gói gọn thông tin hữu ích cho việc phân loại.
* Điều này có những lợi ích sau:
  + [CLS] token được khởi tạo ngẫu nhiên và cập nhật liên tục khi mạng được huấn luyện. Nó có thể mã hóa các đặc điểm thống kê của toàn bộ tập dữ liệu
  + [CLS] token tổng hợp thông tin trên tất cả các tokens khác (tổng hợp đặc trưng toàn cầu), và bởi vì nó không dựa trên chính nội dung hình ảnh nên nó có thể tránh được sự thiên vị đối với một token cụ thể trong chuỗi
  + Sử dụng encoding position cố định cho [CLS] token có thể tránh được đầu ra bị nhiễu bởi encoding position. Trong ViT, tác giả đặt [CLS] token ở vị trí là 0. Bằng cách này, ngay cả khi độ dài của chuỗi thay đổi, encoding position của class embedding vẫn cố định.
* (2), (3): Transformer encoder bao gồm các layer với các khối multiheaded self-attention (MSA) và MLP.
* Layernorm (LN) được áp dụng trước mỗi khối và các residual connections sau mỗi khối. MLP chứa hai layer và một activation function GELU.



* (4): Output của Transformer encoder (z0L) đóng vai trò là biểu diễn hình ảnh y.
* Cả trong quá trình pre-training và fine-tuning, classification head đều được kèm với z0L. Classification head được triển khai bởi MLP với một lớp ẩn ở thời điểm
* pre-training và bởi một lớp tuyến tính duy nhất ở thời điểm fine-tuning.



* Position embeddings được thêm vào patch embeddings để giữ lại thông tin vị trí. Tác giả sử dụng standard learnable 1D position embeddings.
* Các mô hình “Base” và “Large” được áp dụng trực tiếp từ BERT và mô hình “Huge” lớn hơn được thêm vào.



* ViT-L/16 có nghĩa là biến thể “Large” với kích thước input patch 16×16. Lưu ý rằng độ dài chuỗi của Transformer tỷ lệ nghịch với bình phương của kích thước patch và các mô hình có kích thước patch nhỏ hơn sẽ đắt hơn về mặt tính toán.
  1. **Inductive Bias**
* Vision Transformer có độ lệch quy nạp dành riêng cho hình ảnh ít hơn nhiều so với CNN.
* ViT, hay chính xác hơn, các model thuần Attention, thì có inductive bias yếu nên thường phải sử dụng cực cực nhiều data để model có kết quả tốt. Vì vậy, CNN thường mang lại kết quả tốt hơn ViT khi được train với tập dataset nhỏ.
* Với CNN, hay lớp Convolution, ta giả định là những pixel ở gần nhau thì sẽ có liên quan đến nhau

→ đây là Inductive bias đầu tiên của CNN, gọi là tính locality.

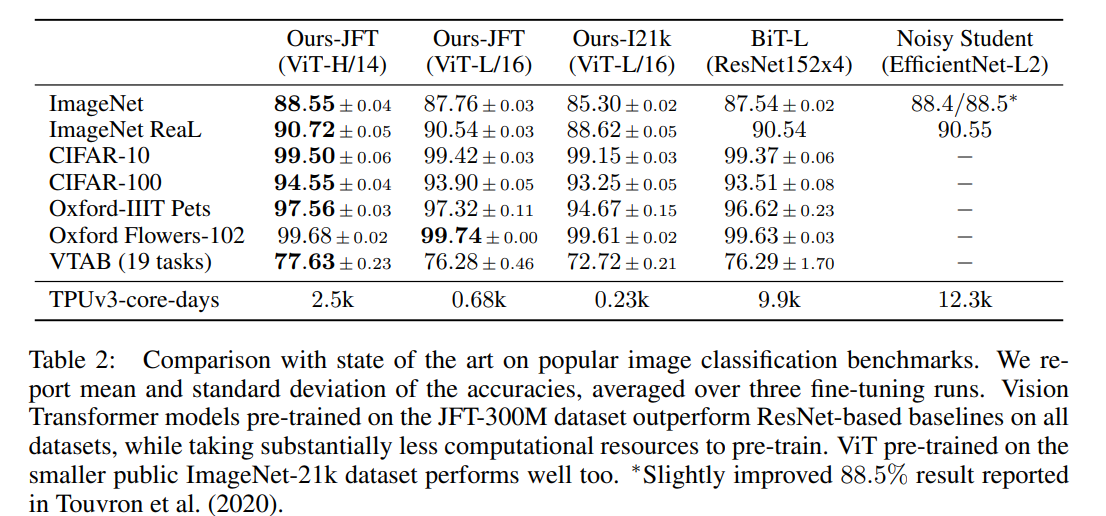
* CNN còn một Inductive bias nữa, gọi là Weight Sharing: Các vùng trên ảnh sẽ được xử lý như nhau, đều được xử lý chung bởi một kernel, nhờ đó, CNN có thể nắm được đặc trưng vật dù vật đó có ở đâu trên ảnh

→ đây là Inductive bias thứ 2 của CNN: weight sharing

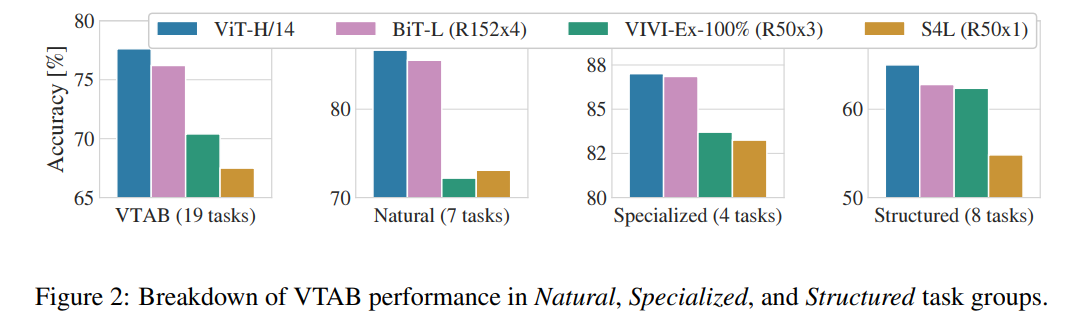
* 1. **Hybrid Architecture**
* Để thay thế cho các raw image patches, chuỗi đầu vào có thể được hình thành từ các feature maps của CNN.
* Trong mô hình kết hợp này, phép chiếu patch embedding E (1) được áp dụng cho các patch được trích xuất từ feature maps CNN.
* Trong trường hợp đặc biệt, các patch có thể có kích thước không gian 1x1, có nghĩa là chuỗi đầu vào thu được bằng cách làm phẳng các kích thước không gian của feature map và chiếu tới kích thước Transformer.
* Classification input embedding and position embeddings được thêm vào như mô tả ở trên.

1. **Experimental & Results**
   1. **Dataset**

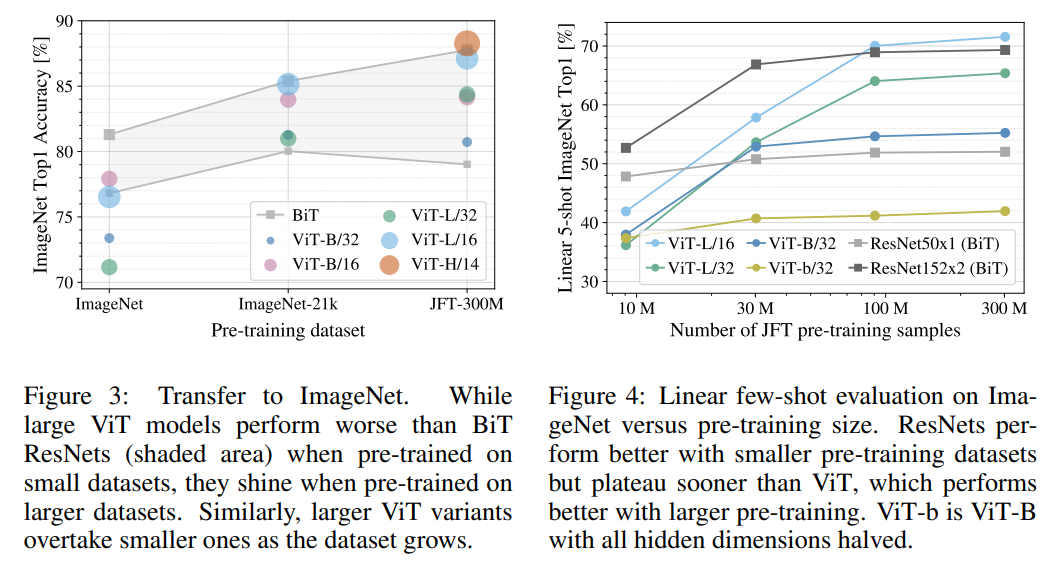
* ImageNet ILSVRC-2012 với 1k class và 1,3 triệu hình ảnh, superset ImageNet-21k với 21k class và 14M hình ảnh và JFT với 18k class và gồm 303M hình ảnh có độ phân giải cao.
  1. **Some Training Details**
* Thông thường, ViT được đào tạo trước trên các tập dữ liệu lớn và được tinh chỉnh cho các tác vụ tiếp theo (nhỏ hơn). Đối với điều này, đầu dự đoán được đào tạo trước sẽ bị loại bỏ và feedforward layer D×K khởi tạo bằng 0 được đính kèm, trong đó K là số lớp của tác vụ nhỏ hơn.
* Việc tinh chỉnh ở độ phân giải cao hơn so với đào tạo trước thường có lợi. Khi cung cấp hình ảnh có độ phân giải cao hơn, kích thước patch được giữ nguyên, dẫn đến độ dài chuỗi hiệu quả lớn hơn.
* Vision Transformer có thể xử lý độ dài chuỗi tùy ý (tối đa hạn chế về bộ nhớ), tuy nhiên, pre-trained position embeddings có thể không còn ý nghĩa nữa. Do đó, phép nội suy 2D của các pre-trained position embeddings được thực hiện, tùy theo vị trí của chúng trong ảnh gốc. (Việc điều chỉnh độ phân giải và trích xuất patch này là những điểm duy nhất mà tại đó inductive bias về cấu trúc 2D của hình ảnh được đưa thủ công vào Vision Transformer)
  1. **SOTA Comparison**



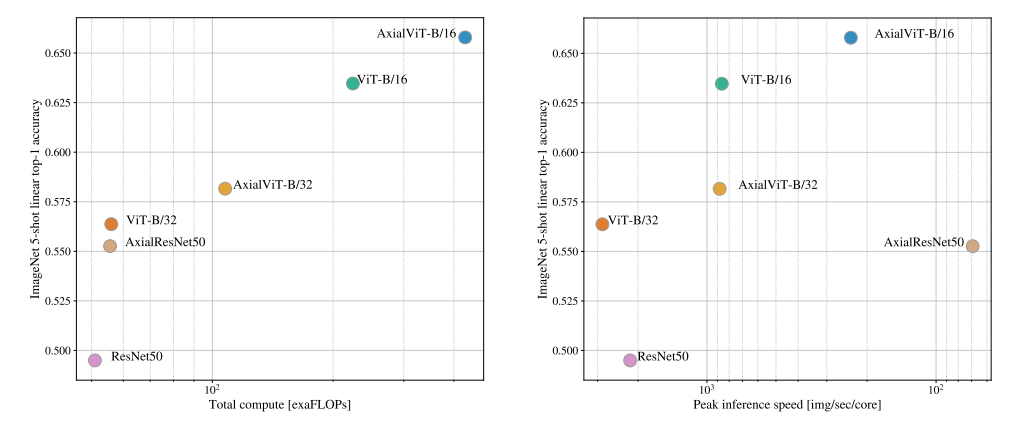
* Mô hình ViT-L/16 nhỏ hơn được huấn luyện trước trên JFT-300M hoạt động tốt hơn BiT-L trong tất cả các nhiệm vụ, đồng thời yêu cầu ít tài nguyên tính toán hơn để huấn luyện.
* ViT-H/14, cải thiện hiệu suất hơn nữa, đặc biệt là trên các bộ dữ liệu khó khăn hơn.



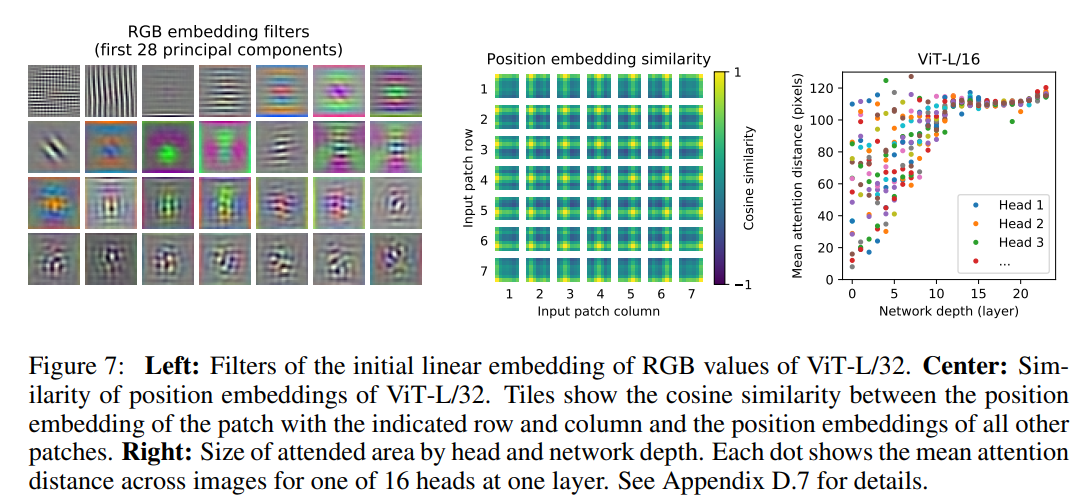
* 1. **Pretraining Data Requirement**

****

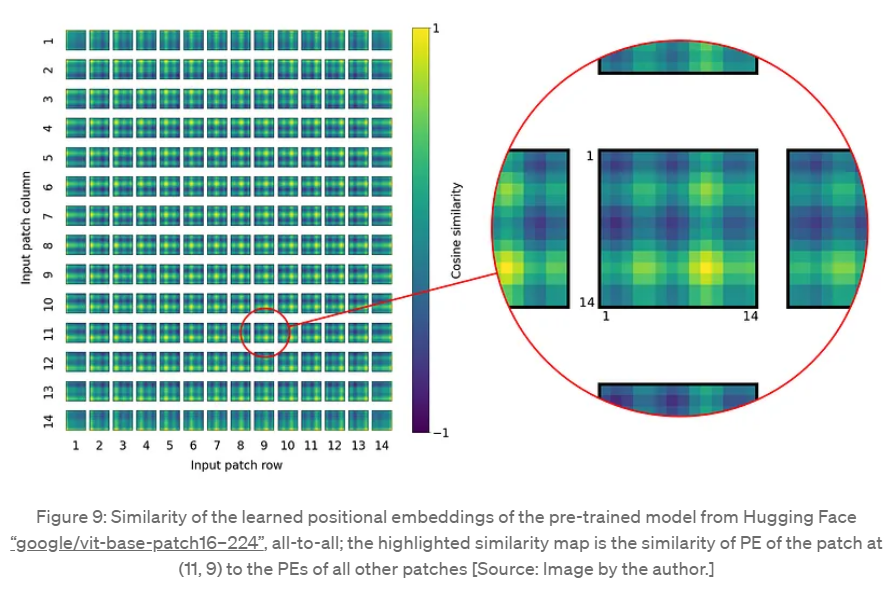
* Trái: Khi được đào tạo trước trên tập dữ liệu nhỏ nhất, các mô hình ImageNet, ViT-Large hoạt động kém hơn so với các mô hình ViT-Base, mặc dù đã được chính quy hóa.
* Với quá trình đào tạo trước ImageNet-21k, hiệu suất của chúng tương tự nhau. Chỉ với JFT-300M, chúng ta mới có thể thấy được toàn bộ lợi ích của các model lớn hơn
* BiT CNN hoạt động tốt hơn ViT trên ImageNet, nhưng với bộ dữ liệu lớn hơn, ViT đã vượt qua
* Phải: Các tập hợp con ngẫu nhiên 9M, 30M và 90M cũng như bộ dữ liệu JFT-300M đầy đủ để huấn luyện trước
* ViT-B/32 nhanh hơn ResNet50 một chút; nó hoạt động kém hơn nhiều trên tập hợp con 9M, nhưng tốt hơn trên các tập hợp con 90M+
* Điều tương tự cũng đúng với ResNet152×2 và ViT-L/16
  1. **Scaling Study**

****

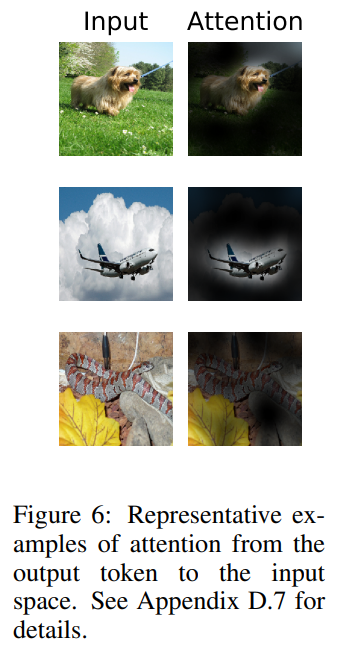
* Vision Transformers áp đảo ResNets về sự cân bằng hiệu suất/tính toán. ViT sử dụng ít tính toán hơn khoảng 2-4× để đạt được hiệu suất tương tự (trung bình trên 5 bộ dữ liệu).
* Hybrids vượt trội hơn ViT một chút với ngân sách tính toán nhỏ, nhưng sự khác biệt sẽ biến mất đối với các mô hình lớn hơn. Kết quả này có phần đáng ngạc nhiên, vì người ta có thể mong đợi việc xử lý tính năng cục bộ tích chập sẽ hỗ trợ ViT ở mọi quy mô.
* Vision Transformers dường như không bão hòa trong phạm vi đã thử
  1. **Inspecting Vision Transformer**

****

* Trái: Các thành phần giống với các hàm cơ sở hợp lý để biểu diễn cấu trúc ở cấp độ thấp trong mỗi patch.
* Giữa: Mô hình học cách mã hóa khoảng cách trong hình ảnh theo độ tương tự của các position embeddings, tức là các patch gần hơn có xu hướng có nhiều position embeddings tương tự hơn.
* Phải: Khoảng cách trung bình trong không gian hình ảnh mà thông tin được tích hợp, dựa trên trọng số chú ý. “Khoảng cách chú ý” này tương tự như kích thước trường tiếp nhận trong CNN.
* Một số heads quan tâm đến hầu hết hình ảnh đã có ở các lớp thấp nhất, cho thấy khả năng tích hợp thông tin trên toàn cầu thực sự đã được mô hình sử dụng.
* Mỗi phần nhúng vị trí tương đồng nhất với chính nó, điều này là hiển nhiên vì nó đại diện cho một vị trí cụ thể trong chuỗi.
* Tuy nhiên, nó cũng rất tương đồng với các phần nhúng lân cận, cho thấy các khu vực gần nhau được mã hóa với các vector tương đồng.
* Hành vi này là mong muốn vì nó cho thấy mô hình đang hiệu quả trong việc capture mối quan hệ không gian giữa các vị trí.
* Mỗi phần nhúng vị trí không chỉ tương tự với các vùng lân cận trực tiếp, mà còn tương tự cao với các vùng hoặc vị trí khác nằm xa hơn. Điều này cho thấy mô hình tổng quát hóa các đặc trưng không chỉ trong khu vực gần mà còn cho các vùng rộng hơn trong chuỗi.



* Mô hình tập trung vào các vùng hình ảnh có liên quan về mặt ngữ nghĩa để phân loại.
* Cơ chế attention tăng cường các phần quan trọng của dữ liệu đầu vào và làm mờ phần còn lại. Lấy ví dụ về khi chú thích một hình ảnh, ta sẽ phải tập trung vào phần có liên quan của hình ảnh để tạo chú thích có ý nghĩa.



* 1. **Self-Supervision**

Với self-supervised pre-training, smaller ViT-B/16 model đạt được độ chính xác 79,9% trên ImageNet, cải thiện đáng kể 2% so với đào tạo từ đầu, nhưng vẫn kém 4% so với đào tạo trước có giám sát.

1. **Note**

* **Ưu điểm của Vision Transformer**:
  + ViT có thể tìm hiểu các đặc trưng chung của hình ảnh. Điều này là do nó có thể chú ý đến bất kỳ phần nào của hình ảnh, bất kể vị trí của nó. Điều này có thể hữu ích cho các tác vụ như object detection và scene understanding.
  + ViT không nhạy cảm với việc tăng cường dữ liệu như CNN. Điều này có nghĩa là nó có thể được đào tạo trên các tập dữ liệu nhỏ hơn.
  + ViT có thể được sử dụng cho nhiều nhiệm vụ phân loại hình ảnh. Điều này bao gồm các nhiệm vụ như object detection, scene understanding
* **Nhược điểm**:
  + Việc đào tạo ViT tốn kém về mặt tính toán [bất kỳ tác vụ nào liên quan đến hình ảnh luôn tốn kém do kích thước pixel của hình ảnh lớn]. Điều này là do chúng có số lượng lớn các tham số.
  + ViT không hiệu quả bằng CNN trong việc xử lý hình ảnh. Điều này là do ViT cần phải chú ý đến mọi phần của hình ảnh, ngay cả khi nó không quan trọng đối với nhiệm vụ hiện tại.
  + ViT không dễ hiểu như CNN. Điều này có nghĩa là rất khó hiểu cách mô hình đưa ra dự đoán
* **Ứng dụng**:
  + Image Classification
  + Image captioning
  + Image segmentation
  + Anomaly detection
  + Action recognition
  + Autonomous driving

**Một vài điều về CNN với ViT:**

* Mặc dù CNN đã được chứng minh trong các tác vụ tầm nhìn máy tính khác nhau và xử lý các bộ dữ liệu quy mô lớn một cách hiệu quả, Vision Transformers mang đến lợi thế trong các tình huống mà sự phụ thuộc toàn cầu và hiểu biết theo ngữ cảnh là rất quan trọng.
* Tuy nhiên, Vision Transformers thường yêu cầu lượng dữ liệu đào tạo lớn hơn để đạt được hiệu suất tương đương với CNN. Ngoài ra, CNN có hiệu quả tính toán do tính chất song song của chúng, làm cho chúng thực tế hơn cho các ứng dụng thời gian thực và hạn chế về tài nguyên.

-------------------Ending Mệt vãi ---------------------