Report paper: A Simple Single-Scale Vision Transformer for Object Detection and Instance Segmentation

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

A Simple Single-Scale Vision Transformer for Object Detection and Instance Segmentation

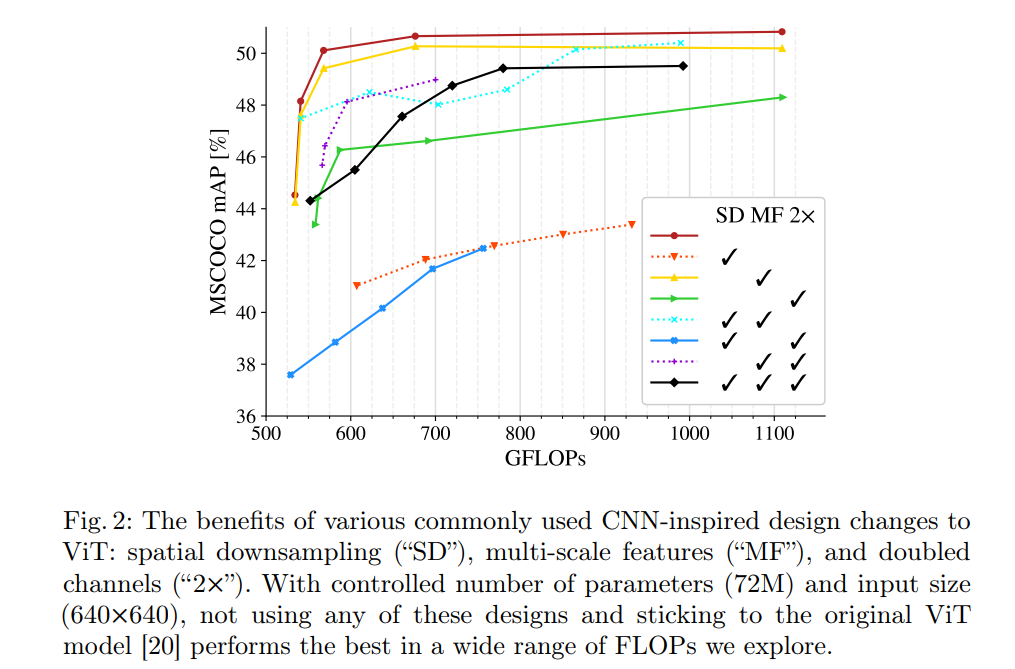
1. **Ý tưởng**

Để áp dụng ViT cho các nhiệm vụ phát hiện đối tượng và phân đoạn, nhiều công trình kế thừa thiết kế nhiều tầng từ mạng tích chập và kiến ​​trúc ViT tùy biến cao. Tuy nhiên, các công trình hiện tại áp dụng thiết kế kiến ​​trúc nhiều tầng như một giải pháp hộp đen mà không hiểu rõ về lợi ích thực sự của nó. Tác giả thực hiện các thử nghiệm chứng minh các giả thuyết của mình và đề xuất một kiến ​​trúc ViT đơn giản và nhỏ gọn có tên là Universal Vision Transformer (UViT) giúp đạt được hiệu suất mạnh mẽ trên điểm chuẩn phân đoạn đối tượng và phát hiện đối tượng COCO. Mục tiêu của tác giả không phải để theo đuổi hiệu suất hiện đại mà là nghiên cứu một cách có hệ thống các nguyên tắc trong thiết kế kiến ​​trúc ViT.

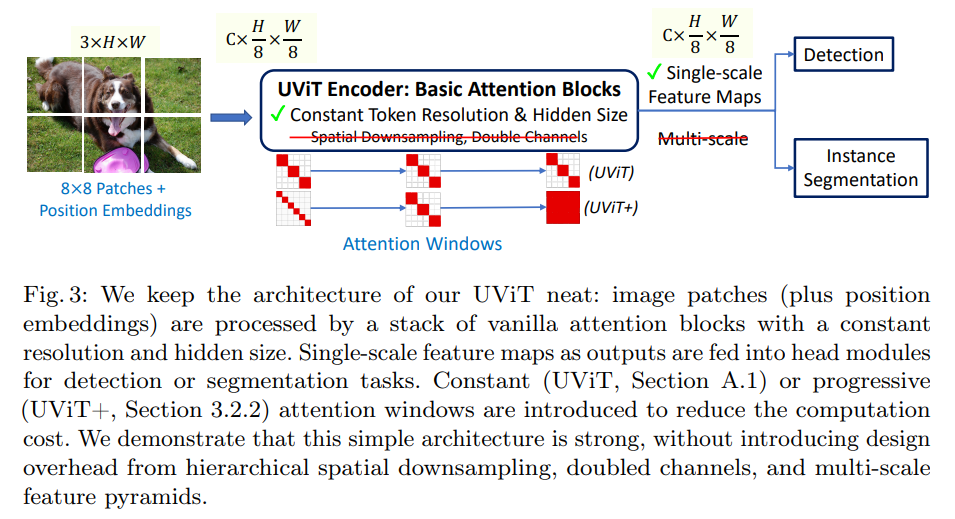
* Những đóng góp của tác giả:
  + Nghiên cứu một cách có hệ thống lợi ích của các thiết kế kiến ​​trúc cơ bản dành cho ViT đối với dense prediction tasks và đề xuất một thiết kế UViT đơn giản cho thấy hiệu suất mạnh mẽ mà không cần thực hiện thủ công các quy ước thiết kế feature pyramid giống CNN vào transformers.
  + Khám phá quy tắc chia tỷ lệ kết hợp mới (độ sâu, chiều rộng, kích thước đầu vào) cho UViT trong dense vision task. Tác giả nhận thấy kích thước đầu vào lớn hơn sẽ tạo ra nhiều không gian hơn để cải thiện thông qua việc mở rộng quy mô mô hình và độ sâu vừa phải (number of attention blocks) hoạt động tốt hơn các blocks nông hơn hoặc sâu hơn.
  + Giảm chi phí tính toán chỉ thông qua attention windows. Tác giả nhận thấy rằng trường tiếp nhận của attention bị hạn chế ở các lớp đầu và attention cục bộ nhỏ gọn là đủ, trong khi chỉ các lớp sâu hơn mới cần đến sự chú ý toàn cầu.
  + Các thử nghiệm về object detection và instance segmentation COCO chứng minh rằng UViT tuy đơn giản nhưng là cơ sở vững chắc cho transformers thực hiện dense prediction tasks.

1. **Phương pháp**
   1. **Is a Simple ViT Design All You Need?**

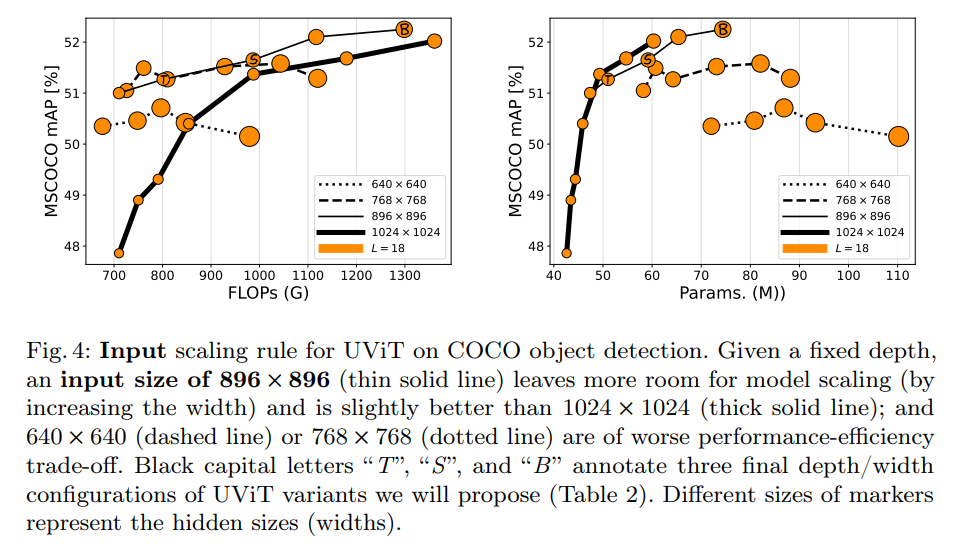
* Tác giả thực hiện một nghiên cứu toàn diện về sự đóng góp của các quy ước thiết kế giống CNN cho ViT trong dense prediction tasks.
* Triển khai: Tác giả sử dụng Cascade Mask-RCNN tiêu chuẩn, với kích thước đầu vào cố định là 640 × 640 trên bộ dữ liệu COCO 2017. Tất cả các mô hình phát hiện đều được tinh chỉnh từ quá trình khởi tạo được huấn luyện trước của ImageNet.
* Cài đặt: Tác giả thực hiện 3 kĩ thuật chính trong thiết kế CNN cho ViT:
  + Spatial downsampling
  + Multi-scale features
  + Doubled channels



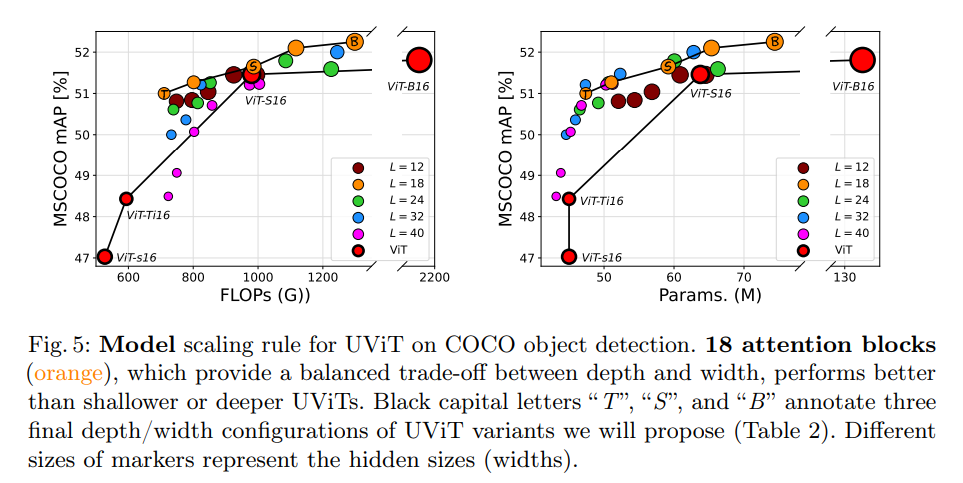
* Kết luận sau nghiên cứu: Tác giả không tìm thấy những lợi ích to lớn khi áp dụng các quy ước thiết kế giống như CNN. Thay vào đó, kiến ​​trúc đơn giản có độ phân giải đặc trưng không đổi và kích thước ẩn có thể là đường cơ sở ViT mạnh mẽ.
  1. **UViT: a Simple Yet Effective Solution**

****

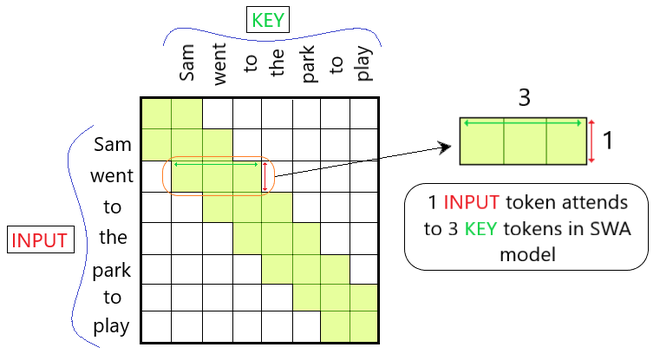
* Lấy các patchs đầu vào 8 × 8, mô hình tìm hiểu cách biểu diễn bằng cách sử dụng token không đổi có độ phân giải tỷ lệ 1/8 (số lượng token vẫn giữ nguyên) và hidden size không đổi (số kênh sẽ không tăng). Một single- scale feature map sẽ được đưa vào detection hoặc segmentation head. Trong khi đó, attention windown sẽ được tận dụng để giảm chi phí tính toán.
* Động lực đằng sau thiết kế UViT không phải là “thêm” bất kỳ lớp nào vào kiến ​​trúc ViT mà thay vào đó là chọn “không thêm” các thiết kế phức tạp. Kiến ​​trúc ViT thông thường cộng với sự cân bằng độ sâu và chiều rộng tốt hơn có thể đạt được hiệu suất cao.
  1. **A Compound Scaling Rule of UViTs**
* Để đạt được sự cân bằng giữa hiệu suất và hiệu quả tốt nhất, tác giả nghiên cứu một cách có hệ thống compounding scaling của UViT theo ba chiều: kích thước đầu vào, độ sâu và chiều rộng. Đối với tất cả các mô hình (điểm đánh dấu vòng tròn), trước tiên tác giả huấn luyện chúng trên ImageNet-1k, sau đó tinh chỉnh chúng trong tác vụ phát hiện COCO.
* Độ sâu (number of attention blocks): các mô hình UViT khác nhau có độ sâu được chọn từ {12, 18, 24, 32, 40}.
* Kích thước đầu vào: ba mức kích thước đầu vào: 640 × 640, 768 × 768, 896 × 896 và 1024 × 1024.
* Chiều rộng (hidden size, or output dimension of attention blocks): điều chỉnh độ rộng thành kiểm soát hơn nữa các kích thước mô hình và chi phí tính toán khác nhau để làm cho các quy tắc chia tỷ lệ khác nhau có thể so sánh được một cách tương đối.



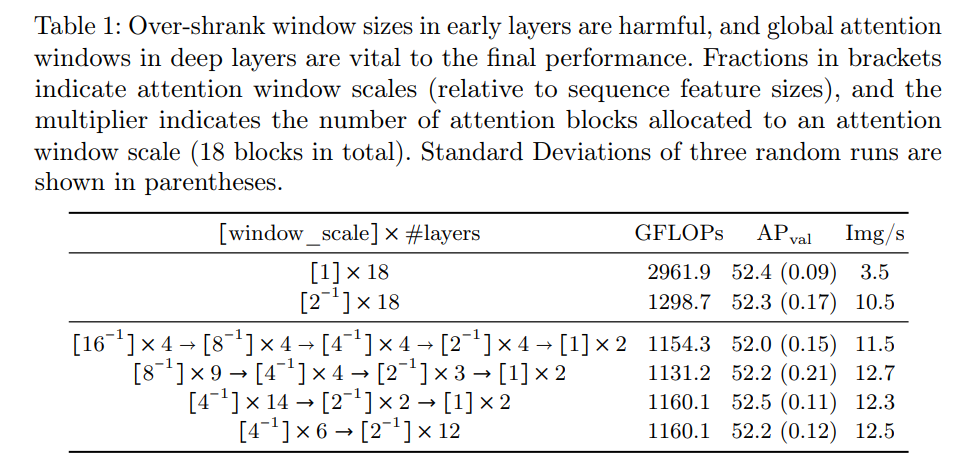
* Nhìn chung, UViT có thể đạt được mAP mạnh với chi phí tính toán (FLOP) vừa phải và số lượng tham số rất nhỏ gọn (thậm chí ít hơn 70M bao gồm cả Cascaded FPN head).
* Đầu vào lớn thường tạo nhiều không gian hơn cho các mô hình mở rộng quy mô hơn nữa. Việc chia tỷ lệ theo kích thước đầu vào 896 × 896 liên tục vượt trội hơn các kích thước đầu vào nhỏ hơn (dẫn đến tình trạng quá khớp mô hình nghiêm trọng) và cũng tốt hơn 1024 × 1024 trong phạm vi FLOP tương đương.
* Đối với độ sâu mô hình (Hình 5), độ sâu khác nhau về màu sắc): khi xem xét cả FLOP và số lượng tham số, 18 khối đạt được hiệu suất tốt hơn các khối 12/24/32/40. Điều này cho thấy UViT cần cân bằng giữa chiều sâu và chiều rộng
* Từ đó, tác giả đề xuất phiên bản UViT cơ bản của mình dưới dạng 18 khối attentions dưới kích thước đầu vào 896 × 896.



* 1. **Attention Windows: a Progressive Strategy**
* Self-attention là một hoạt động toàn cầu. Nhưng trong UViT, attention heads trong các lớp attention ban đầu hoạt động giống như một toán tử cục bộ trong khi ở các lớp sâu hơn, do trường tiếp nhận tăng lên, các attention heads hoạt động giống như một toán tử toàn cục. Do đó, ta có thể giới hạn phạm vi attention của các lớp đầu để giảm chi phí tính toán mà không làm giảm hiệu suất.
* Chiến lược progressive attention window có thể giảm chi phí tính toán của UViT trong khi vẫn duy trì hoặc thậm chí tăng hiệu suất

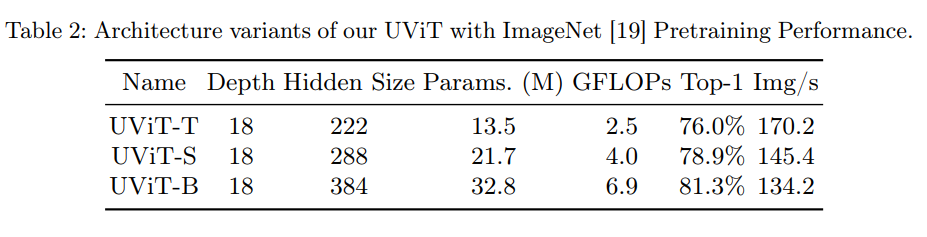


* Từ thử nghiệm của mình, tác giả quan sát được:
  + Với window scale không đổi nhỏ hơn ( 1/2 1/3 1/4 ), tiết kiệm được nhiều chi phí tính toán hơn với sự hy sinh nhỏ trong mAP.
  + Việc áp dụng global attention liên tục trong toàn bộ khối encoder (window size là 1, hàng đầu tiên) phần lớn là dư thừa, điều này mang lại lợi ích cận biên nhưng lại phải chịu chi phí tính toán rất lớn.
  + Early attentions có thể sử dụng các cửa sổ nhỏ hơn như tỷ lệ 1/4, nhưng kích thước cửa sổ thu nhỏ quá mức ( 1/16 , 1/8 ) có thể làm giảm khả năng self-attention (hàng thứ 3, thứ 4).
  + Các lớp sâu hơn vẫn cần có global attention để duy trì hiệu suất cuối cùng (hai hàng cuối cùng).
  + Chiến lược cửa sổ được thiết kế phù hợp (hàng thứ 5) có thể hoạt động tốt hơn các giải pháp thông thường (hàng 1, 2) với chi phí tính toán giảm.

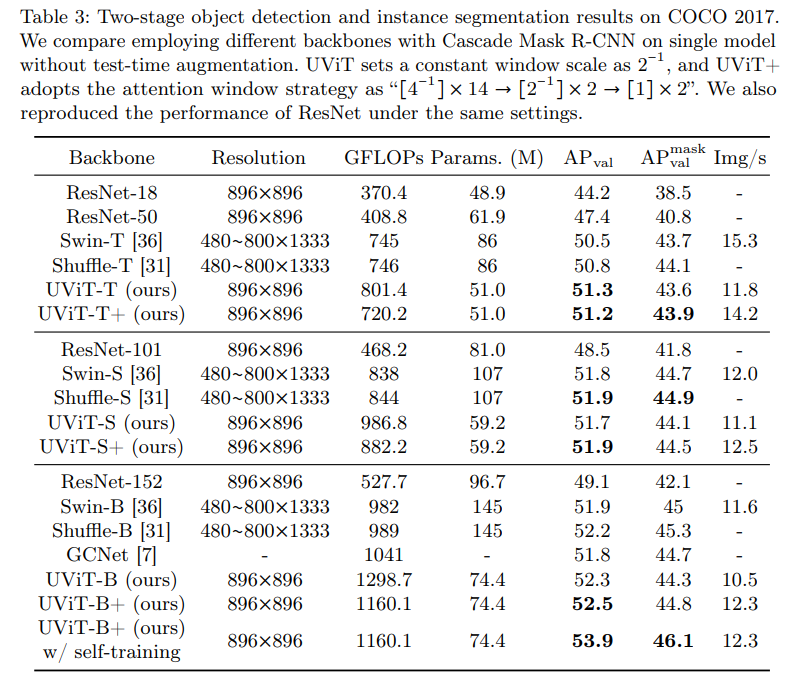


1. **Experimental & Results**
   1. **Implementations**

* Các thí nghiệm được tiến hành trên TPU. Trước khi tinh chỉnh khả năng object detection và instance segmentation, tác giả tuân theo cài đặt đào tạo DeiT để huấn luyện trước UViT trên ImageNet-1k với input size 224 × 224 và batch size 1024.
* Trong quá trình đào tạo trước ImageNet, kernel size của layer linear projection đầu tiên là 16 x 16. Trong quá trình tinh chỉnh, tác giả sử dụng patch size 8 x 8 chi tiết hơn cho mục đích lấy mẫu dày đặc.
* Kernel weight của layer linear projection đầu tiên sẽ được nội suy từ 16 × 16 sang 8 × 8 và position embedding cũng sẽ điều chỉnh bằng phép nội suy.
* Chúng tôi cũng báo cáo thông lượng (“Img/s”), đo lường độ trễ của UViT bằng cách đưa một hình ảnh trên mỗi lõi TPU.
  1. **Architectures and ImageNet Pretraining**
* Số lượng head được cố định là 6 và tỷ lệ mở rộng của mỗi lớp FFN (feed-forward networ) được cố định là 4 trong tất cả các thử nghiệm.



* 1. **COCO detection & instance segmentation**



* Thử nghiệm object detection được thực hiện trên COCO 2017 dataset, chứa 118K training và 5K validation images.
* So với cả CNN và các công trình ViT khác, UViT đạt được kết quả mạnh mẽ với hiệu quả tốt hơn nhiều: với các GFLOP tương tự, UViT sử dụng số lượng tham số ít hơn nhiều (giảm tham số ít nhất 44,9% so với Swin).
* Ngoài ra, tác giả còn tự huấn luyện UViT-B để đánh giá hiệu quả đạt được bằng cách tận dụng unlable data. Tác giả dùng ImageNet-1K không có label như là unlabel set, và mô hình pre-train UViT-B như là teacher model để tạo ra pseudo-label. Tất cả dự đoán với confidence lớn hơn 0.5 được giữ lại cùng với mask tương ứng của nó.
* Đối với tự huấn luyên UViT-B, student model được khởi tạo cùng weight với teacher model. Tỉ lệ label data và pseudo-label data. Ngoài việc tăng training step lên 2x với mỗi epoch thì các hyperparameter không đổi

1. **Conclusion**

* Nghiên cứu chi tiết của tác giả cho thấy rằng các thiết kế giống CNN không phải là điều kiện tiên quyết cho ViT
* Kiến ​​trúc ViT thông thường cộng với quy tắc chia tỷ lệ tốt hơn (độ sâu, chiều rộng, kích thước đầu vào) và progressive attention widow strategy thực sự có thể đạt được hiệu suất phát hiện cao.
* Kiến trúc UViT được đề xuất đạt được hiệu suất mạnh mẽ trên cả COCO object detection và instance segmentation
* Thiết kế của tác giả có tiềm năng hỗ trợ các vấn đề multi-modal/multi-task learning và vision-language.

1. **Note**