

Bảng 1: So sánh các phương pháp tiếp cận tính toán nghiêng về các vị trí khác nhau sử dụng Swin-T. Dấu (\*) là để đề cập tới độ chính xác top-1 trên tập ImageNet-1k được huấn luyện từ đầu. Các mô hình trong cột \* sẽ được sử dụng để đánh giá trên các nhiệm vụ phân loại ảnh của tập ImageNet-1K sử dụng độ phân giải của cửa sổ/ảnh lớn hơn, đánh dấu bởi dấu +. Với các kết quả này, chúng tôi ghi lại cả các kết quả có và không có fine-tuning. Các mô hình này cũng được sử dụng cho tinh chỉnh trên các bài toán phát hiện vật thể trên tập COCO và bài toán phân chia ngữ nghĩa (semantic segmentation) ADE20K.

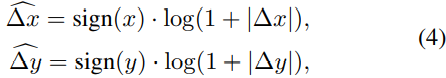
**Continuous relative position bias**: Thay vì trực tiếp tối ưu các bias được tham số hóa, phương pháp tiếp cận continuous relative position bias sử dụng một mạng nhỏ trên các tọa độ tương đối:



Với G là một mạng nhỏ, ví dụ một mạng MLP 2 lớp với một hàm kích hoạt ReLU mặc định ở giữa.

Mạng G sinh ra các giá trị bias cho các tọa độ tương đối tùy ý, và vì vậy có thể được chuyến tiếp một cách tự nhiên để tinh chỉnh các bài toán với kích thước cửa sổ thay đổi tùy ý. Trong khi suy diễn, các giá trị bias ở mỗi vị trí tương đối có thể được tính toán trước và lưu lại như các tham số của mô hình, như vậy suy diễn cũng giống như cách tiếp cận bias được tham số hóa ban đầu.

**Các tọa độ không gian logarit**: Khi chuyển đổi theo các kích thước cửa sổ thay đổi lớn, một phần lớn của khoảng tọa độ tương đối cần được ngoại suy. Để giảm nhẹ vấn đề này, chúng tôi đề xuất sử dụng các tọa độ không gian logarit thay vì không gian tuyến tính như lúc đầu:



Với deltax, deltay và deltaxhat ,deltayhat là các tọa độ tương ứng ở không gian log và được scale tuyến tính.

Bằng việc sử dụng các tọa độ không gian log (log-spaced), khi ta chuyển đổi các bias vị trí tương đối theo các độ phân giải của cửa sổ, tỉ lệ ngoại suy yêu cầu sẽ nhỏ hơn bias mà sử dụng tọa độ không gian tuyến tính ban đầu. Lấy ví dụ chuyển đổi từ kích thước cửa sổ 8x8 được tiền huấn luyện sang một kích thước cửa sổ 16x16 đã được tinh chỉnh, sử dụng các tọa độ gốc ban đầu, khoảng tọa độ đầu vào sẽ là từ [-7,7]x[-7,7] tới [-15,15]x[-15,15]. Tỉ lệ ngoại suy là khoảng ban đầu. Sử dụng các tọa độ không gian log, khoảng đầu vào sẽ từ tới . Tỉ lệ ngoại suy là khoảng ban đầu, tức là nhỏ hơn khoảng 4 lần tỉ lệ ngoại suy sử dụng tọa độ không gian tuyến tính ban đầu.

Bảng 1 so sánh hiệu năng chuyển đổi của các hướng tính toán khác nhau thiên về vị trí. Có thể thấy rằng hướng tiếp cận thiên vị trí liên tục không gian log (log-spaced CPB) cho hiệu năng tốt nhất, đặc biệt khi được chuyển đổi sang các kích thước cửa sổ lớn hơn.

3.4. Tiền huấn luyện tự giám sát (self-supervised pre-training):

Các mô hình lớn hơn thì cũng cần nhiều dữ liệu hơn. Để giải quyết vấn đề cần nhiều dữ liệu, các mô hình thị giác lớn trước đây điển hình khai thác các tập dữ liệu có nhãn lớn nhưn JFT-3B. Trong bài báo này, chúng tôi khai thác phương pháp tiền huấn luyện tự giám sát, SimMIM, để giải quyết yêu cầu cần dữ liệu có nhãn. Bằng cách này, chúng tôi đã huấn luyện thành công một mô hình Swin Transformer rất mạnh với 3 tỷ tham số và đạt được kết quả SOTA trên 4 tập đánh giá đại diện, với chỉ 70 triệu ảnh có nhãn (bằng 1/40 tập JFT-3B).

3.5. Cách triển khai để tiết kiệm bộ nhớ GPU:

Một vấn đề khác nằm ở việc sử dụng bộ nhớ GPU khá tốn kém với cách triển khai thông thường khi cả năng lực và độ phân giải đều lớn. Để giải quyết vấn đề về bộ nhớ, chúng tôi áp dụng các cách triển khai sau:

* Hàm tối ưu không dư (Zero-redundancy Optimizer)(ZeRO). Trong cách triển khai các hàm tối ưu song song hóa dữ liệu, các tham số mô hình và trạng thái tối ưu được truyền trực tiếp tới mỗi GPU. Cách triển khai này rất ko phù hợp trong việc sử dụng bộ nhớ GPU, ví dụ, một mô hình với 3 tỉ tham số sẽ tiêu tốn 46GB bộ nhớ GPU khi một hàm tối ưu AdamW và tham số/trạng thái fp32 được sử dụng. Với một hàm tối ưu ZeRO, các tham số của mô hình và các trạng thái tối ưu tương ứng sẽ được phân chia và phân bố tới nhiều GPUs đồng thời, điều nãy sẽ giảm đáng kể việc sử dụng bộ nhớ GPU. Chúng tôi áp dụng DeepSpeed framework và sử dụng tùy chọn ZeRO stage-1 trong các thí nghiệm. Việc tối ưu này ít có ảnh hưởng tới tốc độ huấn luyện.
* Điểm kiểm tra sự kích hoạt (Activation check-pointing). Feature maps trong các lớp của Transformer cũng tiêu tốn khá nhiều bộ nhớ GPU, điều này có thể gây ra nghẽn cổ chai khi ảnh và độ phân giải của cửa sổ đều lớn. Kỹ thuật kiểm tra sự kích hoạt có thể giảm đáng kể tiêu tốn bộ nhớ, trong khi tốc độ huấn luyện có thể chậm hơn khoảng 30%.
* Tính toán tự tập trung tuần tự (sequential self-attention computation). Để huấn luyện các mô hình lớn trên độ phân giải rất lớn, ví dụ, một bức ảnh độ phân giải với kích thước cửa sổ , với GPU A100 100GB bộ nhớ cũng vẫn không đủ, thâm chi với cả 2 kỹ thuật tối ưu vừa trình bày ở trên. Chúng tôi nhận thấy rằng trong trường hợp này, mô đun tự tập trung (self-attention) gây ra sự nghẽn cổ chai. Để giảm nhẹ vấn đề này, chúng tôi triển khai tính toán tự tập trung một cách tuần tự, thay vì sử dụng cách tiếp cận tính toán theo tệp (batch) như trước đây. Cách tối ưu này được áp dụng cho các lớp trong 2 tầng đầu tiên và ít có ảnh hưởng tới tốc độ huấn luyện.

Với các cách triển khai này, chúng tôi có thể huấn luyện một mô hình 3 tỷ tham số sử dụng GPU A100-40GB của NVIDIA cho bài toán phát hiện vật thể trên tập COCO với độ phân giải đầu vào là và bài toán phân loại hành động Kinetics-400 với độ phân giải đầu vào là .

3.6. Các cấu hình cho mô hình:

Chúng tôi giữ nguyên thiết lập các tầng, khối và các kênh của mô hình Swin Transformer gốc với 4 thiết lập của Swin Transformer V2:

* Swin V2-T: C = 96, #. Khối = {2,2,6,2}
* Swin V2-S/B/L: C=96/128/192, #. Khối = {2,2,18,2}

Với C là số kênh trong tầng đầu tiên.

Chúng tôi tiếp tục tăng kích thước của Swin Transformer V2 lên cỡ lớn và cỡ rất lớn, tương ứng với 658 triệu tham số và 3 tỷ tham số:

* SwinV2-H: C = 352, #. Khối = {2, 2, 18,2}
* SwinV2-G: C = 512, #. Khối = {2,2,42,4}

Với SwinV2-H và SwinV2-G, chúng tôi thêm một lớp được chuẩn hóa trên nhánh chính sau mỗi 6 lớp. Để tiết kiệm thời gian thí nghiệm, chúng tôi chi tiến hành với SwinV2-G cho các thí nghiệm kích cỡ lớn. SwinV2-H được tiến hành cho các nghiên cứu song song về học tự giám sát.

4. Các thí nghiệm:

4.1. Các bài toán và tập dữ liệu:

Chúng tôi tiến hành các thí nghiệm trên tập phân loại ảnh ImageNet-1K (V1 và V2), tập phát hiện vật thể COCO, và tập phân loại ngữ nghĩa ADE20K. Với các thí nghiệm của mô hình 3 tỷ tham số, chugns tôi cũng thống kê về độ chính xác trên tập nhận diện hành động trên video Kinetics-400.

* Bài toán phân loại ảnh. Tập dữ liệu ImageNet-1K V1 và V2 được sử dụng để đánh giá. ImageNet-22K gồm có 14 triệu ảnh và 22 nghìn loại đối tượng được tùy chọn tiến hành cho tiền huấn luyện. Với mô hình SwinV2-G lớn nhất được tiền huấn luyện, một tập dữ liệu mở rộng ImageNet-22K được thu thập riêng với hơn 70 triệu ảnh đã được sử dụng. Với tập dữ liệu này, một quy trình loại bỏ ảnh trùng lặp dã được tiến hành để loại bỏ các bức ảnh chồng lấn nhau trên tập validation ImageNet-1K V1 và V2.
* Bài toán phát hiện vật thể. COCO được sử dụng để đánh giá. Với các thí nghiệm cho mô hình lớn nhất, chúng tôi tiến hành một bước tiền huấn luyện việc phát hiện vật thể sử dụng tập dữ liệu Object 365 v2, ở giữa bước tiền huấn luyện cho phân loại hình ảnh và bước tinh chỉnh trên tập COCO.
* Semantic segmentation. Tập dữ liệu ADE20K được sử dụng.
* Phân loại hành động từ video. Tập Kinetics-400 (K400) được sử dụng để đánh giá.

Các thiết lập về tinh chỉnh và tiền huấn luyện sẽ được đề cập chi tiết trong phần phụ lục.

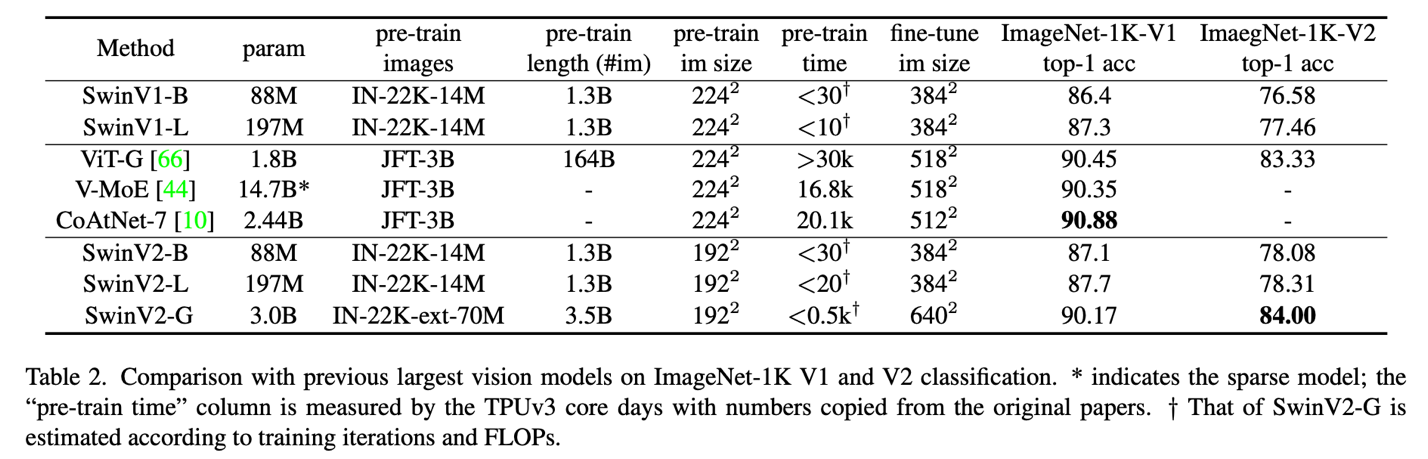
4.2. Mở rộng các thí nghiệm:

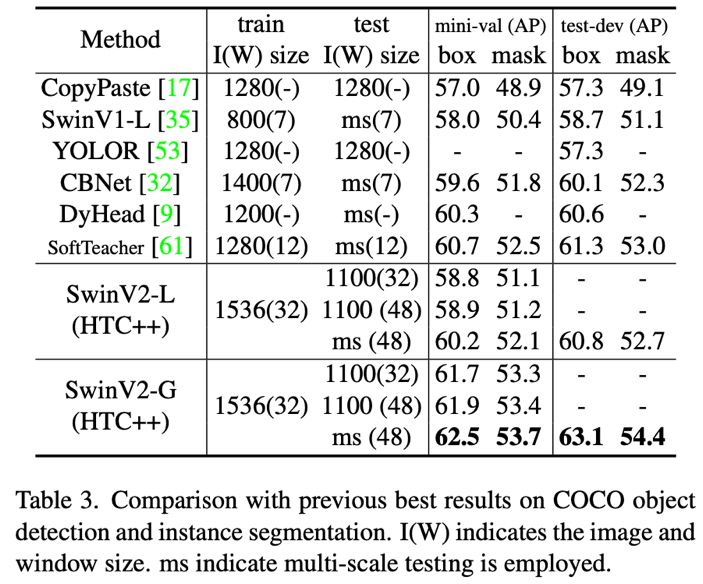
Chúng tôi trước hết trình bày về các kết quả trên các đánh giá trực quan bằng việc tăng kích thước của mô hình tới 3 tỷ tham số và độ phân giải cao cho cửa sổ/ảnh.

Các thiết lập cho thí nghiệm với mô hình SwinV2-G. Chúng tôi sử dụng độ phân giải nhỏ hơn cho ảnh với 192x192 pixel trong bước tiền huấn luyện để tiết kiệm chi phí huấn luyện. Chúng tôi tiến hành bước tiền huấn luyện theo 2 bước. Đầu tiên, mô hình được tiền huấn luyện sử dụng một phương pháp tự giám sát trên tập dữ liệu mở rộng ImageNet-22K-ext với 20 epochs. Sau đó, mô hình được tiền huấn luyện với 30 epochs sử dụng bài toán phân loại hình ảnh trên tập dữ liệu này. Các thiết lập về tinh chỉnh và tiền huấn luyện sẽ được mô tả trong phần phụ lục.

Trong các phần tiếp theo, chúng tôi trình bày về độ chính xác của SwinV2-G trên đánh giá thị giác đại diện (representative vision benchmarks). Lưu ý rằng vì mục đích chính của chúng tôi là để tìm ra làm thế nào để dễ dàng mở rộng năng lực của mô hình và độ phân giải cửa sổ, và liệu các bài toán về thị giác có được lợi ích từ năng lực lớn hơn này không, chúng tôi không so sánh riêng về độ phức tạp hoặc dữ liệu tiền huấn luyện ở đây.

Các kết quả của bài toán phân loại hình ảnh trên tập ImageNet-1K. Bảng 2 so sánh SwinV2-G với các mô hình thị giác tốt nhất trước đây trên ImageNet-1K V1 và V2. SwinV2-G là mô hình thị giác lớn nhất trong số này. Nó đạt độ chính xác top-1 84% trên đánh giá ImageNet V2, 0.7% cao hơn mô hình tốt nhất trước đây (83.3%). Chúng tôi đạt độ chính chỉ thấp hơn một chút trên tập ImageNet-1K V1 (90.17% so với 90.88%). Sự khác biệt có thể đến từ mức độ over tuning khác nhau trên tập dữ liệu. Lưu ý rằng, chúng tôi cũng sử dụng ít vòng lặp huấn luyện hơn nhiều và độ phân giải cũng thấp hơn so với các mô hình trước đây, trong khi vẫn cho kết quả rất tốt.



Table

Description automatically generated

Table

Description automatically generated

Chúng tôi cũng tiến hành so sánh SwinV2-B và SwinV2-L với mô hình gốc SwinV1-B và SwinV1-L, và thấy tương ứng có sự cải thiện lần lượt 0.8% và 0.4%. Sự cải thiện nhỏ hơn ở SwinV2-L so với SwinV2-B ám chỉ rằng nếu vượt quá kích thước này, thì cần nhiều dữ liệu hơn, regularization mạnh hơn, hoặc các phương pháp học tự giám sát cao cấp hơn cần được áp dụng.

Kết quả phát hiện vật thể trên tập COCO. Bảng 3 so sánh SwinV2-G với các kết quả tốt nhất trước đây trên bài toán phát hiện vật thể và instance segmentation của tập dữ liệu COCO. Nó đạt kết quả 63.1/54.4 box/max AP trên tập test-dev của COCO, với tương ứng +1.8/1.4 cao hơn các kết quả tốt nhất trước đây (61.3/53.0). Điều này chỉ ra rằng việc tăng kích cỡ của mô hình thị giác là có ích cho các bài toán nhận diện thị giác hoặc phát hiện vật thể. Hướng tiếp cận của chúng tôi có thể sử dụng kích thước cửa sổ khác để kiểm tra các ích lợi tăng thêm, có thể đóng góp một cách hiệu quả tới hướng tiếp cận CPB không gian Log.

Các kết quả semantic segmentation trên tập ADE20K. Bảng 4 so sánh SwinV2-G với các kết quả tốt nhất trước đây trên tập đánh giá ADE20K. Mô hình đã đạt được độ chính xác 59.9 mIoU trên tập val ADE20K, cao hơn 1.5 so với kết quả tốt nhất trước đây (58.4). Điều này chỉ ra rằng việc tăng kích thước mô hình thị giác là có ích cho các bài toán nhận diện thị giác mức độ pixel. Sử dụng kích thước cửa sổ lớn hơn lúc đánh giá có thể tăng thêm 0.2 độ chính xác, có thể đóng góp tới hướng tiếp cận CPB không gian Log.

Các kết quả phân loại hành động từ video trên tập Kinetics-400. Bảng 5 so sánh SwinV2-G với các mô hình tốt nhất trước đây trên tập đánh giá phân loại hành động Kinetics-400. Mô hình đạt được độ chính xác 86.8% top-1 accuracy, +1.4% cao hơn kết quả tốt nhất. Điều này chỉ ra rằng việc tăng kích thước mô hình thị giác cũng đem lại ích lợi cho các bài toán nhận diện video. Trong ngữ cảnh này, sử dụng kích thước cửa sổ lớn hơn lúc đánh giá có thể đem lại +0.2% độ chính xác, có thể đóng góp tới hướng tiếp cận CPB không gian Log.

Table

Description automatically generated

Table

Description automatically generated

4.3. Các nghiên cứu về việc lược bỏ:

Lược bỏ phần về cụm các lớp res-post-norm và scaled cosin attention: Bảng 6 lược bỏ phần hiệu năng cho đề xuất cụm các lớp res-post-norm và scaled cosin attention cho mô hình Swin Transformer. Cả 2 kỹ thuật đều cải thiện độ chính xác cho các kích thước rất nhỏ, nhỏ và gốc, tương ứng là +0.2%, +0.4% và +0.5%, thêm vào đó, các kỹ thuật này cũng có ích lợi với các mô hình lớn hơn. Ví dụ như nó cũng có ích với kiến trúc ViT (+0.4%). Hướng chuẩn hóa được đề xuất cũng tốt hơn một số cách chuẩn hóa khác, như được trình bày trong Bảng 7.

Quan trọng hơn, việc kết hợp post-norm và scaled cosine attention giúp ổn định quá trình huấn luyện. Như được trình bày trong bảng 2, trong khi các giá trị kích hoạt ở các lớp sâu hơn ở mô hình Swin Transformer gốc trở nên rất lớn ở kích thước lớn, mô hình phiên bản mới này cho thấy tác động nhẹ hơn nhiều. Với mô hình kích thước rất lớn, việc tiền huấn luyện tự giám sát bị phân kỳ khi sử dụng mô hình Swin Transformer gốc, trong khi huấn luyện tốt trên mô hình Swin Transformer V2.

Tăng độ phân giải cửa sổ bằng các hướng tiếp cận khác: Bảng 1 và 8 loại bỏ phần hiệu năng của 3 cách tiếp cận bằng cách tăng độ phân giải cửa sổ từ 256x256 lúc tiền huấn luyện tới các kích thước lớn hơn trong 3 bài toán phân loại hình ảnh trên tập ImageNet-1K, nhận diện vật thể COCO, và semantic segmentation ADE20K. Có thể thấy rằng:

* 1) Các hướng tiếp cận khác nhau cho kết quả tương tự nhau ở bước tiền huấn luyện (81.7%-81.8%)
* 2) Khi chuyển đổi tới các bài toán down-stream, hướng tiếp cận bias vị trí tương đối liên tục (CPB) thể hiện tốt hơn hướng tiếp cận bias được tham số hóa sử dụng trong Swin Transformer V1. So sánh với cách tiếp cận linear-spaced, phiên bản log-spaced tốt hơn một chút.
* 3) Thay đổi càng lớn về độ phân giải giữa tiền huấn luyện và bước tinh chỉnh, thì hướng tiếp cận CPB log-spaced cho lợi ích càng lớn.

Table

Description automatically generated

Trong Bảng 1 và Bảng 8, chúng tôi cũng trình bày về độ chính xác sử dụng độ phân giải cửa sổ đích mà không cần tinh chỉnh (fine-tuning) (xem con số đầu tiên trong mỗi cột ở thí nghiệm tập ImageNet-1K). Độ chính xác nhận diện vẫn không kém ngay cả khi kích thước cửa sổ tăng từ 8 tới 24 (78.9% so với 81.8%), trong khi độ chính xác top-1 accuracy của hướng tiếp cận ban đầu giảm đáng kể từ 81.7% xuống 68.7%. Cũng lưu ý rằng khi không có tinh chỉnh, sử dụng kích thước cửa sổ 12 mà mô hình tiền huấn luyện chưa từng thấy có thể đem lại độ chính xác tăng thêm +0.4% so với độ chính xác ban đầu. Điều này chỉ ra rằng chúng ta có thể cải thiện độ chính xác thông qua việc điều chỉnh cửa sổ lúc đánh giá, như được quan sát trong Bảng 3,4, và 5.

5. Kết luận:

Chúng tôi và trình bày các kỹ thuật cho việc tăng kích thước của mô hình Swin Transformer lên tới 3 tỷ tham số và giúp mô hình có khả năng huấn luyện được với kích cỡ ảnh lên tới 1,536 x 1,536 pixel, bao gồm các kỹ thuật res-post-norm và scaled cosine attention để giúp dễ dàng hơn tăng năng lực của mô hình, cũng như hướng tiếp cận thiên lệch vi trí tương đối liên tục không gian log để giúp mô hình hiệu quả hơn khi chuyển đổi độ phân giải của cửa sổ. Kiến trúc mới này được đặt tên là Swin Transformer V2, và bằng việc tăng năng lực và độ phân giải của mô hình, nó đã đạt kỷ lục mới về độ chính xác trên cả 4 đánh giá đại diện cho các bài toán về thị giác. Nhờ những kết quả mạnh mẽ này, chúng tôi kỳ vọng sẽ tiến hành nhiều nghiên cứu nữa theo hướng này để có thể tiệm cận được năng lực xử lý của các mô hình ngôn ngữ và tạo điều kiện cho việc mô hình hóa liên hợp giữa 2 lĩnh vực thị giác và ngôn ngữ.

Lời cảm ơn:

Chúng tôi cảm ơn các đồng nghiệp ở Microsoft nói chung, và Eric Chang, Lidong Zhou, Jing Tao, Aaron Zhang, Edward Cui, Bin Xiao, Lu Yuan, Peng Cheng, Fan Yang nói riêng về những thảo luận hữu ích và sự giúp đỡ về tài nguyên GPU cũng như dữ liệu.

Tham khảo

[1]  Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucˇic ́, and Cordelia Schmid. Vivit: A video vision transformer, 2021. 3, 7

[2]  Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 3

[3]  HangboBao,LiDong,andFuruWei.Beit:Bertpre-training of image transformers, 2021. 2, 7

[4]  Andrew Brock, Soham De, and Samuel L Smith. Character- izing signal propagation to close the performance gap in unnormalized resnets. *arXiv preprint arXiv:2101.08692*, 2021. 4

[5]  Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Sub- biah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakan- tan, Pranav Shyam, Girish Sastry, Amanda Askell, Sand- hini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Rad- ford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1, 2

[6]  TianqiChen,BingXu,ChiyuanZhang,andCarlosGuestrin. Training deep nets with sublinear memory cost, 2016. 2, 6

[7]  Bowen Cheng, Alexander G. Schwing, and Alexander Kir- illov. Per-pixel classification is not all you need for semantic segmentation. *arXiv*, 2021. 7

[8]  Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haib- ing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers, 2021. 3

[9]  Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions, 2021. 7

[10]  Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes, 2021. 2, 3, 5, 7

[11]  Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[12]  Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[13]  Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text- to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021. 3, 8

[14]  Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows, 2021. 3

[15]  Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl- vain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representa- tions*, 2021. 2, 3

[16] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2021. 1, 2

[17] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung- Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2020. 7

[18] PriyaGoyal,MathildeCaron,BenjaminLefaudeux,MinXu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchin- sky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised pretraining of visual features in the wild, 2021. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[20] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceed- ings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018. 3

[21] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3464–3473, October 2019. 3

[22] Shihua Huang, Zhichao Lu, Ran Cheng, and Cheng He. Fapn: Feature-aligned pyramid network for dense image prediction, 2021. 7

[23] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer, 2021. 3

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. 3

[25] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. 2

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics hu- man action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[27] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training, 2021. 3

[28] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8, 2019. 2

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing sys- tems*, pages 1097–1105, 2012. 2

[30]  Yann LeCun, Le ́on Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[31]  Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision trans- formers, 2021. 3

[32]  Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cb- netv2: A composite backbone network architecture for ob- ject detection, 2021. 7

[33]  Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dolla ́r, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[34] Ze Liu, Han Hu, Yue Cao, Zheng Zhang, and Xin Tong. A closer look at local aggregation operators in point cloud analysis, 2020. 3

[35]  Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin trans- former: Hierarchical vision transformer using shifted win- dows, 2021. 1, 2, 3, 5, 7

[36]  Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021. 2,7

[37]  Microsoft. Turing-nlg: A 17-billion-parameter language model by microsoft, 2020. 1, 2

[38] Microsoft. Using deepspeed and megatron to train megatron-turing nlg 530b, the world’s largest and most powerful generative language model, 2021. 1, 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 6

[40] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsuper-  
vised multitask learners. 2019. 1, 2, 3

[41]  Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learn- ing Research*, 21(140):1–67, 2020. 1, 2, 3

[42]  Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020. 2, 5

[43]  Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to im- agenet?, 2019. 2, 6, 7

[44] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, Andre ́ Susano Pinto, Daniel  
Keysers, and Neil Houlsby. Scaling vision with sparse mix-ture of experts, 2021. 2, 3, 5, 7

[45] Michael S. Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos?, 2021. 2, 7, 8

[46] Kristof T Schu ̈tt, Pieter-Jan Kindermans, Huziel E Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Mu ̈ller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv preprint arXiv:1706.08566*, 2017. 3

[47] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 6

[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, May 2015. 1, 2

[49] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve ́ Je ́gou. Training data-efficient image transformers & distillation through at- tention. *arXiv preprint arXiv:2012.12877*, 2020. 3

[51] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. In- stance normalization: The missing ingredient for fast styliza- tion, 2017. 3

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko- reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2, 3,8

[53] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for mul- tiple tasks, 2021. 7

[54] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. Deep parametric continu- ous convolutional neural networks. *2018 IEEE/CVF Confer- ence on Computer Vision and Pattern Recognition*, Jun 2018. 3

[55] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyra- mid vision transformer: A versatile backbone for dense pre- diction without convolutions, 2021. 3

[56] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer, 2021. 3

[57] Yuxin Wu and Kaiming He. Group normalization. In *Pro- ceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 3

[58] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dolla ́r, and Ross Girshick. Early convolutions help trans- formers see better. *arXiv preprint arXiv:2106.14881*, 2021. 3

[59] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A sim- ple framework for masked image modeling. In *Tech report*, 2022.4,5,6,8

[60]  Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the trans- former architecture. 2020. 3

[61]  Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to- end semi-supervised object detection with soft teacher, 2021. 2, 7

[62]  JianweiYang,ChunyuanLi,PengchuanZhang,XiyangDai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers, 2021. 3

[63]  Zhuliang Yao, Yue Cao, Yutong Lin, Ze Liu, Zheng Zhang, and Han Hu. Leveraging batch normalization for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 413–422, 2021. 4

[64]  Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet, 2021. 3

[65]  Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recog- nition. *arXiv preprint arXiv:2106.13112*, 2021. 3

[66]  XiaohuaZhai,AlexanderKolesnikov,NeilHoulsby,andLu- cas Beyer. Scaling vision transformers, 2021. 2, 3, 5, 7

[67]  Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision long- former: A new vision transformer for high-resolution image encoding, 2021. 3

[68]  Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmen- tation from a sequence-to-sequence perspective with trans- formers. *arXiv preprint arXiv:2012.15840*, 2020. 3

[69]  Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fi- dler, Adela Barriuso, and Antonio Torralba. Semantic under- standing of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018. 6

Transformer => Vit Transformer => Swin Transformer => Swin Transformer V2

<https://www.youtube.com/watch?v=iDulhoQ2pro&t=885s&ab_channel=YannicKilcher>

<https://www.youtube.com/watch?v=TrdevFK_am4&t=1381s&ab_channel=YannicKilcher>

<https://www.youtube.com/watch?v=SndHALawoag&ab_channel=AICoffeeBreakwithLetitia>

<https://www.youtube.com/watch?v=aircAruvnKk&t=1032s&ab_channel=3Blue1Brown>