*Tóm tắt*

Chúng tôi trình bày các kỹ thuật mở rộng Swin Transformer lên đến 3 tỷ thông số và làm cho nó có khả năng huấn luyện những hình ảnh có độ phân giải lên đến 1.536 × 1.536. Bằng cách mở rộng dung lượng và độ phân giải, Swin Transformer thiết lập các kỷ lục mới về bốn tiêu chuẩn thị giác biểu diễn: độ chính xác top-1 phân loại hình ảnh ImageNet-V2 là 84,0%, phát hiện đối tượng COCO là 63,1/54,4 box/mask mAP, phân đoạn ngữ nghĩa ADE20K là 59,9 mIoU, và độ chính xác top 1 phân loại hoạt động trong video Kinetics-400 là 86,8%.

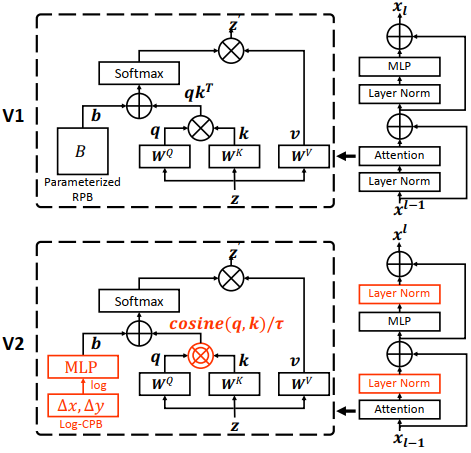
Chúng tôi giải quyết tính không ổn định trong huấn luyện mô hình và nghiên cứu cách chuyển hiệu quả các mô hình được tiền huấn luyện có độ phân giải thấp sang các mô hình có độ phân giải cao hơn. Với mục tiêu này, một số công nghệ mới được đề xuất: 1) kỹ thuật phần dư sau chuẩn hóa và phương pháp tiếp cận theo tỷ lệ cosine để cải thiện tính ổn định của các mô hình thị giác lớn; 2) kỹ thuật thiên vị vị trí liên tục khoảng cách giữa các log để chuyển một cách hiệu quả các mô hình tiền huấn luyện với các hình ảnh có độ phân giải thấp sang các hình ảnh có độ phân giải cao hơn. Ngoài ra, chúng tôi cũng chia sẻ các chi tiết quan trọng trong việc triển khai giúp tiết kiệm đáng kể mức tiêu thụ bộ nhớ GPU và do đó, việc huấn luyện các mô hình thị giác lớn bằng GPU thông thường trở nên khả thi. Sử dụng các kỹ thuật này và việc tiền huấn luyện có giám sát, chúng tôi đã huấn luyện thành công mô hình Swin Transformer 3 tỷ thông số và chuyển nó thành nhiều nhiệm vụ thị giác khác với các hình ảnh có độ phân giải cao, đạt được độ chính xác hiện đại theo nhiều tiêu chuẩn.

**1. Giới thiệu**

Việc mở rộng các mô hình ngôn ngữ đã thành công ngoài sức tưởng tượng. Nó cải thiện đáng kể hiệu suất mô hình ngôn ngữ và thể hiện khả năng few-shot ngôn ngữ tương tự như con người một cách đáng kinh ngạc. Kể từ mô hình lớn BERT với 340 triệu tham số, các mô hình ngôn ngữ nhanh chóng được nhân rộng hơn 1.000 lần trong vài năm, đạt tới 530 tỷ tham số đặc và 1.6 nghìn tỷ tham số thưa. Các mô hình ngôn ngữ lớn này cũng được cho là có khả năng ngôn ngữ few-shot ngày càng mạnh mẽ giống như trí thông minh của con người.

Mặt khác, việc mở rộng quy mô của các mô hình thị giác đã bị tụt hậu. Mặc dù từ lâu người ta đã nhận ra rằng các mô hình thị giác lớn hơn thường thực hiện tốt hơn, nhưng kích thước tuyệt đối của các mô hình gần đây nhất mới chỉ có thể đạt khoảng 1-2 tỷ thông số. Quan trọng hơn, không giống như các mô hình ngôn ngữ lớn, các mô hình thị giác lớn hiện có chỉ được áp dụng cho việc phân loại hình ảnh.

Để huấn luyện thành công mô hình thị giác tổng quát và lớn, chúng ta cần giải quyết một số vấn đề chính. Thứ nhất, các thử nghiệm của chúng tôi với các mô hình thị giác lớn cho thấy tính không ổn định trong huấn luyện. Chúng tôi thấy rằng sự khác biệt về biên độ hoạt hóa giữa các lớp trở nên lớn hơn đáng kể trong các mô hình lớn. Xem xét kỹ hơn kiến ​​trúc ban đầu cho thấy điều này là do đầu ra của khối dư được thêm trực tiếp vào nhánh chính. Kết quả là các giá trị kích hoạt được tích lũy theo từng lớp, và do đó biên độ ở các lớp sâu hơn lớn hơn đáng kể so với các biên độ ở các lớp ban đầu. Để giải quyết vấn đề này, chúng tôi đề xuất một cấu hình chuẩn hóa mới, được gọi là res-post-norm, chuyển lớp LN từ phần đầu của đơn vị dư sang phần phụ trợ, như trong Hình 1. Chúng tôi nhận thấy cấu hình mới này tạo ra các giá trị kích hoạt nhẹ hơn nhiều trên các lớp mạng. Chúng tôi cũng đề xuất một scaled cosine attention thay thế dot product attention trước đó. Scaled cosine attention làm cho việc tính toán không liên quan đến biên độ đầu vào khối, và các giá trị attention ít có khả năng rơi vào các giá trị cực trị. Trong các thí nghiệm của chúng tôi, hai kỹ thuật được đề xuất không chỉ giúp quá trình huấn luyện ổn định hơn mà còn cải thiện độ chính xác, đặc biệt là đối với các mô hình lớn hơn.



*Hình 1. Để mở rộng quy mô mô hình và nâng cao độ phân giải, một số điều chỉnh được thực hiện trên kiến trúc Swin Transformer ban đầu (V1): 1) Một res-post-norm* *thay thế cấu hình pre-norm trước đó; 2) Scaled cosine attention thay thế dot product attention ban đầu; 3) Phương pháp tiếp cận chệch vị trí tương đối liên tục log-spaced thay thế cách tiếp cận tham số hóa trước đó. Các Adapter 1) và 2) giúp dễ dàng mở rộng quy mô mô hình. Adapter 3) làm cho mô hình biến chuyển hiệu quả hơn qua các độ phân giả. Kiến trúc này tên là Swin Transformer V2.*

Thứ hai, nhiều tác vụ thị giác xuôi dòng như phát hiện đối tượng và phân đoạn ngữ nghĩa yêu cầu hình ảnh đầu vào có độ phân giải cao hoặc attention lớn. Sự thay đổi khẩu độ tiền huấn luyện độ phân giải thấp và tinh chỉnh độ phân giải cao có thể khá lớn. Phổ biến hiện nay là thực hiện phép nội suy hai khối của các ánh xạ chệch vị (position bias map). Cách khắc phục đơn giản này khá đặc biệt và kết quả thường là dưới tối ưu. Chúng tôi giới thiệu độ chệch vị liên tục log-based (Log-CPB), giúp tạo ra các giá trị chệch vị đối với các khoảng tọa độ tùy ý bằng cách áp dụng một mạng meta nhỏ với đầu vào là tọa độ dạng log-based. Vì mạng meta lấy bất kỳ tọa độ nào, nên một mô hình tiền huấn luyện sẽ có thể tự do chuyển qua các khẩu độ bằng cách chia sẻ trọng số của mạng meta. Một thiết kế quan trọng trong cách tiếp cận của chúng tôi là chuyển đổi các tọa độ dạng log để tỷ lệ ngoại suy có thể thấp ngay cả khi khẩu độ mục tiêu lớn hơn đáng kể so với khẩu độ tiền huấn luyện. Việc mở rộng dung lượng và độ phân giải của mô hình cũng dẫn đến mức tiêu thụ bộ nhớ GPU cao đến báo động so với các mô hình thị giác hiện có. Để giải quyết vấn đề về bộ nhớ, chúng tôi kết hợp một số kỹ thuật quan trọng bao gồm zero-optimizer, trỏ kiểm tra kích hoạt (activation check pointing) và tính toán selfs-attention tuần tự. Với các kỹ thuật này, mức tiêu thụ bộ nhớ GPU của các mô hình lớn và độ phân giải cao giảm đáng kể.

Với các kỹ thuật trên, chúng tôi đã huấn luyện thành công mô hình Swin Transformer với 3 tỷ thông số và chuyển hiệu quả sang các nhiệm vụ thị giác khác nhau với độ phân giải hình ảnh lớn tới 1.536 × 1.536, sử dụng GPU Nvidia A100-40G. Trong xử lý mô hình tiền huấn luyện, chúng tôi cũng sử dụng tiền huấn luyện tự giám sát để giảm sự phụ thuộc vào dữ liệu lớn gắn nhãn. Với dữ liệu được dán nhãn ít hơn 40 lần so với dữ liệu thực tế trước đây (JFT-3B), mô hình với 3 tỷ thông số đạt được độ chính xác hiện đại trên một loạt các tiêu chuẩn thị giác. Cụ thể, nó đạt được độ chính xác top 1phân loại hình ảnh ImageNet-V2 là 84,0 phát hiện đối tượng COCO là 63,1/54,4 box/mask mAP, phân đoạn ngữ nghĩa ADE20K là 59,9 mIoU, và độ chính xác top 1 phân loại hoạt động trong video Kinetics-400 là 86,8%, cao hơn + NA%, + 4,4 / + 3,3, +6,3 và +1,9 so với các con số tốt nhất trong Swin Transformers ban đầu và vượt qua kỷ lụ trước đó + 0,8%, + 1,8 / + 1,4, +1,5 và + 1,4%.

Bằng cách mở rộng cả quy mô và độ phân giải của các mô hình thị giác hiệu suất cao đối với các nhiệm vụ thị giác chung, giống hiệu suất mô hình xử lý ngôn ngữ tự nhiên NLP chung, chúng tôi mong muốn tiến hành nhiều nghiên cứu theo hướng này hơn để có thể thu hẹp khoảng cách giữa mô hình thị giác và các mô hình ngôn ngữ và tạo điều kiện thuận lợi cho việc tạo mô hình chung của hai lĩnh vực.

**2. Các công trình liên quan**

**Mạng ngôn ngữ và mở rộng** Các công trình tiên phong về Transformer được xem là các mạng tiêu chuẩn. Việc khám phá mở rộng quy mô kiến trúc này đã bắt đầu và được đẩy nhanh nhờ việc phát minh ra các phương pháp học tự giám sát, chẳng hạn như mô hình ngôn ngữ ẩn hay tự động hồi quy, và đã được khuyến khích hơn nữa bởi luật scaling. Kể từ đó, dung lượng của các mô hình ngôn ngữ đã tăng đáng kể hơn 1.000 lần trong vài năm, từ BERT-340M đến Megatron-Turing-530B và Switch-Transformer-1.6T thưa. Với quy mô tăng, độ chính xác các chuẩn ngôn ngữ khác nhau đã được cải thiện đáng kể. Hiệu suất zero-shot hoặc few-shot cũng được cải thiện đáng kể, đây là nền tảng của trí tuệ con người.

**Mạng thị giác và mở rộng** CNN từ lâu đã trở thành mạng thị giác máy tính tiêu chuẩn. Kể từ AlexNet, các kiến trúc ngày càng trở nên sâu hơn và lớn hơn, điều này đã nâng cao rất nhiều tác vụ trực quan khác nhau và phần lớn thúc đẩy làn sóng học sâu trong thị giác máy tính, chẳng hạn như VGG, GoogleNet và ResNet citehe2015resnet. Trong hai năm qua, kiến trúc CNN đã được mở rộng hơn nữa lên khoảng 1 tỷ tham số, tuy nhiên, hiệu suất tuyệt đối có thể không được khuyến khích, có lẽ do thiên vị quy nạp trong kiến trúc CNN đã hạn chế sức mạnh mô hình.

Năm ngoái, Transformers bắt đầu lần lượt đạt được hết các chuẩn hình ảnh, bao gồm chuẩn phân loại cấp ảnh ImageNet-1K, chuẩn phát hiện đối tượng cấp vùng COCO, chuẩn phân đoạn ngữ nghĩa cấp pixel ADE20K, chuẩn phân loại video Kinetics-400, v.v. Kể từ khi những chuẩn này được sử dụng, nhiều biến thể Transformer thị giác đã được cải thiện độ chính xác ở quy mô tương đối nhỏ. Chỉ có một số công trình mở rộng Transformers thị giác. Tuy nhiên, chúng dựa trên một tập lớn dữ liệu ảnh được gán nhãn, hay JFT-3B và chỉ được áp dụng cho các bài toán phân loại ảnh.

**Truyền qua cửa sổ / độ phân giải nhân** Đối với CNN, các công trình trước đây thường cố định kích thước nhân trong quá trình tiền huấn luyện và tinh chỉnh. Các Transformer thị giác toàn cục, như ViT, tính toán attention toàn cục, với khẩu độ attention tỷ lệ thuận với độ phân giải ảnh đầu vào. Đối với kiến trúc Transformer thị giác cục bộ, chẳng hạn Transformer Swin, khẩu độ có thể được cố định hoặc thay đổi trong quá trình tinh lọc. Việc cho phép khẩu độ thay đổi sẽ thuận tiện hơn khi sử dụng, để có thể chia đều toàn bộ ánh xạ đặc tính và điều chỉnh các trường tiếp nhận để có độ chính xác tốt hơn. Để xử lý các khẩu độ thay đổi giữa tiền huấn luyện và tinh chỉnh, nội suy hai khối là phương pháp phổ biến trước đây. Trong bài báo này, chúng tôi đề xuất cách tiếp cận chệch vị liên tục log-spaced (Log-CPB) để chuyển các trọng số mô hình tiền huấn luyện độ phân giải thấp một cách trơn tru thành các cửa sổ có độ phân giải cao hơn.

**Nghiên cứu về các khái niệm lệch** Trong NLP, phương pháp chệch vị tương đối tỏ ra có ích, so với phương pháp nhúng vị tuyệt đối được sử dụng trong Transformer ban đầu. Trong thị giác máy tính, phương pháp chệch vị tương đối được sử dụng phổ biến hơn, có lẽ bởi vì các mối quan hệ không gian của tín hiệu thị giác đóng một vai trò quan trọng hơn trong mô hình trực quan. Một thực tế phổ biến là học trực tiếp các giá trị chệch dưới dạng trọng số của mô hình. Cũng có một số công trình đặc biệt nghiên cứu cách đặt và học các thuật ngữ chệch.

**Tích chập liên tục và các biến thể** Cách tiếp cận Log-CPB của chúng tôi cũng liên quan đến các công trình trước đây về tích chập liên tục và các biến thể, sử dụng mạng meta để xử lý các điểm dữ liệu không đều. Phương pháp Log-CPB của chúng tôi được truyền cảm hứng từ những nỗ lực trong khi giải quyết một vấn đề khác là chuyển các chệch vị tương đối trong Transformer thị giác với các khẩu độ tùy ý. Chúng tôi cũng đề xuất các tọa độ log-spaced để giảm bớt khó khăn trong việc ngoại suy khi chuyển giữa các thay đổi khẩu độ lớn.

3. Swin Transformer V2

**3.1. Khái quát về Swin Transformer**

Swin Transformer là xương sống thị giác máy tính nói chung và đã đạt được mức hiệu suất lớn trong nhận dạng chi tiết như phát hiện đối tượng cấp vùng, phân đoạn ngữ nghĩa cấp pixel và phân loại hình ảnh theo cấp độ hình ảnh. Ý tưởng chính của Swin Transformer là đưa một số yếu tố hình ảnh quan trọng vào bộ mã hóa vani Transformer, bao gồm hệ thống phân cấp, địa phương và bất biến dịch, trong đớ kết hợp sức mạnh của: đơn vị Transformer cơ bản có khả năng mô hình hóa mạnh và các yếu tố tạo ảnh thân thiện.

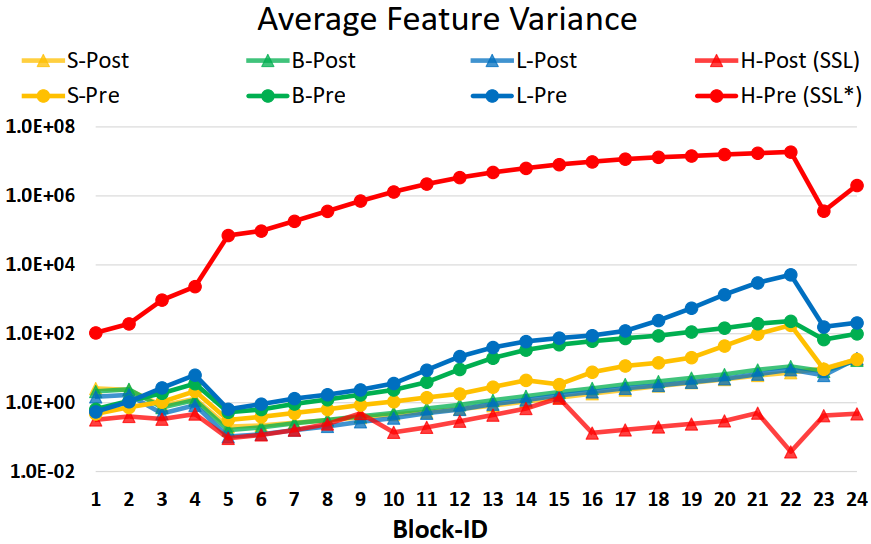
**Cấu hình chuẩn** Các công nghệ chuẩn hóa rất quan trọng trong việc huấn luyện một cách ổn định các kiến trúc sâu. Swin Transformer ban đầu kế thừa các Transformer ngôn ngữ và vani ViT để cấu hình chuẩn hóa trước mà không cần nghiên cứu sâu, như trong hình 1. Trong các phần mục nhỏ sau, chúng ta sẽ xem xét cấu hình chuẩn hóa mặc định này.

**Độ chệch vị tương đối** là một thành phần quan trọng trong Swin Transformer đời đầu, nó đưa ra một thuật ngữ chệch tham số bổ sung để mã hóa mối quan hệ hình học trong tính toán:

Attention(Q, K, V ) = SoftMax(QKT /√d + B)V, (1)

trong đó B ∈ RM^2 × M^2 là số hạng chệch vị tương đối cho mỗi đầu; Q, K, V ∈ RM^2 × d là các ma trận truy vấn, khóa và giá trị; d là tỉ lệ query/key và M2 là số ô trong một khung. Độ chệch vị tương đối mã hóa các cấu hình không gian tương đối của các yếu tố ảnh và được thể hiện trong nhiều tác vụ trực quan, đặc biệt đối với các tác vụ nhận dạng như phát hiện đối tượng.

Trong Transformer Swin, các vị trí tương đối dọc theo mỗi trục nằm trong khoảng [−M + 1, M - 1] và độ chệch vị tương đối được tham số hóa dưới dạng ma trận thiên vị Bˆ ∈ R (2M − 1) × (2M − 1), và các phần tử trong B được lấy từ Bˆ. Khi chuyển qua các khẩu độ khác nhau, ma trận chệch vị tương đối trong khóa tiền huấn luyện được sử dụng để khởi tạo ma trận chệch vị có kích thước khác trong tinh chỉnh bằng nội suy hai khối.



*Hình 2. Biểu đồ lan truyền tín hiệu với các kích thước mô hình khác nhau. Các mô hình kích thước H được huấn luyện ở giai đoạn học tự giám sát và các kích thước khác được huấn luyện bằng nhiệm vụ phân loại ảnh. \* cho thấy mô hình 40-epoch được sử dụng trước khi gặp sự cố.*

**Các vấn đề trong việc mở rộng mô hình và độ phân giải** Chúng tôi quan sát thấy hai vấn đề khi mở rộng quy mô và độ phân giả của Transformer Swin.

• *Một là tính không ổn định khi mở rộng mô hình*. Như trong Hình 2, khi chúng tôi mở rộng mô hình Swin Transformer ban đầu từ kích thước nhỏ sang kích thước lớn, các giá trị kích hoạt ở các lớp sâu hơn sẽ tăng lên đáng kể. Sự khác biệt giữa các lớp có biên độ cao nhất và thấp nhất đã đạt đến giá trị cực đại là 104. Khi chúng tôi mở rộng nó lên một kích thước khổng lồ (658 triệu tham số), nó không thể hoàn thành quá trình huấn luyện, như thể hiện trong Hình 3.

• *Hiệu suất bị giảm sút khi chuyển các mô hình giữa những độ phân giải khác nhau*. Trong hàng đầu tiên Bảng 1, độ chính xác giảm đáng kể khi chúng tôi trực tiếp kiểm tra độ chính xác của mô hình ImageNet-1K được tiền huấn luyện (ảnh 256×256 với khẩu độ 8×8) ở độ phân giải hình ảnh lớn hơn và khẩu độ thông qua phương pháp nội suy hai khối. Có thể cần kiểm tra lại cách tiếp cận chệch vị tương đối trong Swin Transformer gốc.

Trong các mục sau, chúng tôi trình bày các kỹ thuật giải quyết những vấn đề này, bao gồm hậu chuẩn hóa phần dư và cosine attention được nâng cấp để giải quyết vấn đề không ổn định và phương pháp tiếp cận chệch vị liên tục log-spaced để giải quyết việc chuyển giữa các độ phân giải khác nhau.



*Hình 3. SwinV1-H so với SwinV2-H trong huấn luyện.*

**3.2 Mở rộng mô hình**

Như đã đề cập trong Phần 3.1, Transformer Swin ban đầu (và hầu hết các Transformer thị giác) sử dụng một lớp chuẩn lớp ở đầu mỗi khối, kế thừa từ vani ViT. Khi mở rộng mô hình, sự gia tăng đáng kể giá trị kích hoạt được quan sát thấy ở các lớp sâu hơn. Trên thực tế, trong cấu hình tiền chuẩn hóa, các giá trị kích hoạt đầu ra của mỗi khối dư được hợp nhất trực tiếp trở lại nhánh chính, và biên độ của nhánh chính ngày càng lớn hơn ở các lớp sâu hơn. Sự chênh lệch biên độ lớn ở các lớp khác nhau gây ra sự mất ổn định trong quá trình huấn luyện.

**Sau chuẩn hóa** Để giải quyết vấn đề này, chúng tôi đề xuất sử dụng phương pháp chuẩn hóa sau dư, như Hình 1. Trong cách tiếp cận này, đầu ra của mỗi khối dư được chuẩn hóa trước khi hợp nhất trở lại nhánh chính và biên độ của nhánh chính không cộng dồn khi lớp xuống sâu hơn. Như trong Hình 2, biên độ kích hoạt theo cách tiếp cận này nhẹ hơn nhiều so với cấu hình trước khi chuẩn hóa ban đầu.

Trong quá trình huấn luyện mô hình lớn nhất, chúng tôi giới thiệu một lớp chuẩn hóa lớp bổ sung trên nhánh chính cứ sau 6 khối Transformer, để ổn định hơn nữa trong quá trình huấn luyện.

**Scaled cosine attention** Trong tính toán self-attention ban đầu, các số hạng tương tự của các cặp pixel được tính như một tích giữa truy vấn và các vectơ khóa. Chúng tôi thấy rằng khi cách tiếp cận này được sử dụng trong các mô hình trực quan lớn, các ánh xá attention đã học của một số khối và phần đầu thường bị chi phối bởi một vài cặp pixel, đặc biệt là trong cấu hình res-post-standard. Để giải quyết vấn đề này, chúng tôi đề xuất một phương pháp tiếp cận scaled cosine attention tính toán logit attention của một cặp pixel i và j bằng một hàm cosine:

Sim(qi, kj) = cos(qi, kj)/τ + Bij, (2)

trong đó Bij là độ lệch vị trí tương đối giữa pixel i và j; τ là một đại lượng vô hướng có thể học được, phần tử và các lớp riêng biệt. τ được đặt lớn hơn 0,01. Hàm cosine được chuẩn hóa một cách tự nhiên và do đó có thể có các giá trị attention nhẹ hơn.