Swin Transformer Hierarchical Vision Transformer using Shifted Windows

Background

- Transformer가 Vision Task에서는 효과가 적고, NLP 분야에서는 효과적인 이유에 대해 2가지 차이점을 바탕으로 설명한다.

- 첫째, Vision Task에서는 **시각적 객체들이 서로 다른 Scale**을 가진다. 하지만 모델들이 항상 Fixed Scale로 접근하기 때문에 Object Detection이나, Segmentation과 같이 Scale에 민감한 Task를 잘 처리하지 못한다.

- 둘째, Image Segmentation과 같은 경우 픽셀 하나 하나에 대해서 굉장히 민감한 Task인데 ViT 방식의 경우 고해상도 Image에 대해서 Quadratic 하게 증가하는 **연산량으로 인해서 고해상도 이미지를 그대로 사용할 수 없거나, 학습에 굉장히 오랜 시간과 비용이 든다는** 점이다.

Motivation

- Shifted Window 개념을 통해 **다양한 Scale을 살필 수 있도록 계층 구조**를 만들고, 이미지 Size에 대해서 Linear하게 연산량이 증가하는 방식의 Backbone을 제안한다.

- **작은 Patch부터 계산을 진행하며, 점점 병합을 통해 큰 Patch를 확인하는** 계층적 방식을 가지고 있으며 이를 통해 마치 FPN이나 U-Net과 같이 다양한 Object Scale을 고려할 수 있게 된다.

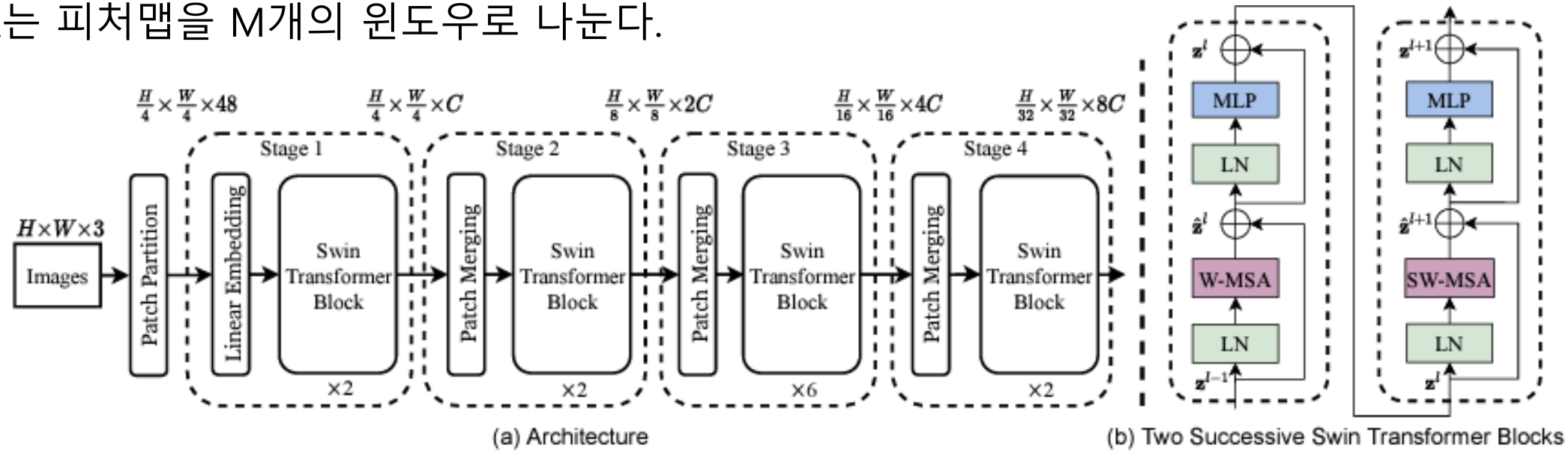
- 또한 **Shifted Window 내부에 존재하는 patch**들 간에만 Self Attention을 계산하는 방식을 통해 계산량을 획기적으로 감소시켰다.

Model

- Patch Partition, Linear Embedding, Swin Transformer Block, Patch Merging으로 구성되어 있다.

- Swin Transformer Block은 2개의 Encoder로 구성되어 있으며, Window Multi-Head Self Attention(W-MSA), Shifted Window Multi-Head Self Attention(SW-MSA)으로 구성된다.

- W-MSA는 피처맵을 M개의 윈도우로 나눈다.



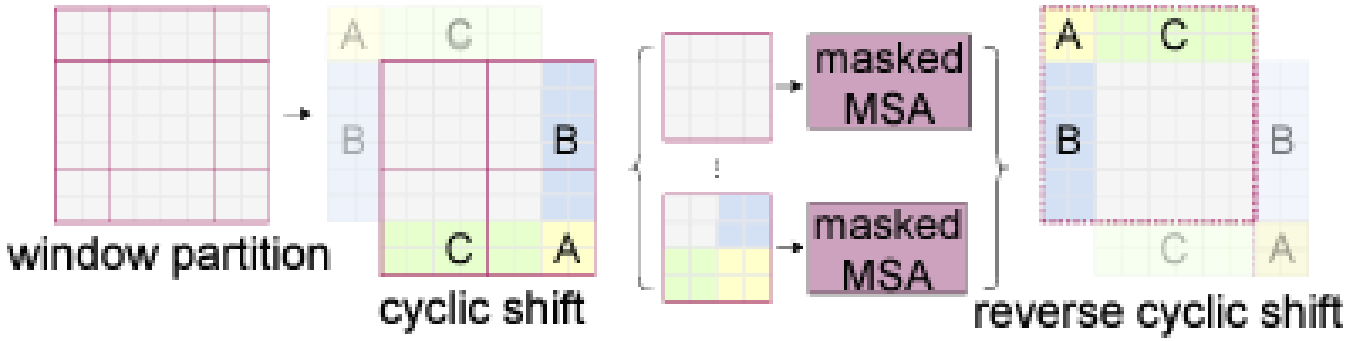
<Vision Transformer 구조>

Model

- W-MSA만을 이용할 경우, Window간의 연관관계를 파악할 수 없고 이로 인해 이미지 전체를 인식하는데 어려움이 발생한다

- Window를 Shift하고 Padding을 추가하는 방식도 가능하고 구현상 더 편리하지만, 그렇게 할 경우 연산량이 크게 증가하는 부작용이 존재한다. 그래서 아래 그림과 같이 Cyclic Shift를 진행한 뒤, 실제로는 이웃하지 않은 Patch인 A,B,C에 대해서는 Mask를 통해 Attention을 계산하지 않도록 한다.

- 그 후, 원상태로 복원시킴으로써, 효율적으로 Window간의 상관관계를 학습할 수 있도록 한다.



<SW-MSA 과정>

Result

- ImageNet과 같은 중간 크기의 데이터셋에 이 모델을 적용하였을 때는 좋지 않은 성능을 보였으나 충분히 큰 스케일에서 vision transformer를 사전 학습한 결과, 더 적은 데이터셋을 가진 하위 태스크에 전이 학습하여 좋은 성능을 얻을 수 있었다.

(a) Regular ImageNet-1K trained models					
method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
RegNetY-4G [48]	224^2	21M	4.0G	1156.7	80.0
RegNetY-8G [48]	224^2	39M	8.0G	591.6	81.7
RegNetY-16G [48]	224^2	84M	16.0G	334.7	82.9
EffNet-B3 [58]	300^2	12M	1.8G	732.1	81.6
EffNet-B4 [58]	380^2	19M	4.2G	349.4	82.9
EffNet-B5 [58]	456^2	30M	9.9G	169.1	83.6
EffNet-B6 [58]	528^2	43M	19.0G	96.9	84.0
EffNet-B7 [58]	600^2	66M	37.0G	55.1	84.3
ViT-B/16 [20]	384^2	86M	55.4G	85.9	77.9
ViT-L/16 [20]	384^2	307M	190.7G	27.3	76.5
DeiT-S [63]	224^2	22M	4.6G	940.4	79.8
DeiT-B [63]	224^2	86M	17.5G	292.3	81.8
DeiT-B [63]	384^2	86M	55.4G	85.9	83.1
Swin-T	224^2	29M	4.5G	755.2	81.3
Swin-S	224^2	50M	8.7G	436.9	83.0
Swin-B	224^2	88M	15.4G	278.1	83.5
Swin-B	384^2	88M	47.0G	84.7	84.5

<사전 학습된 데이터에 따른 benchmark 데이터에 대한 성능>