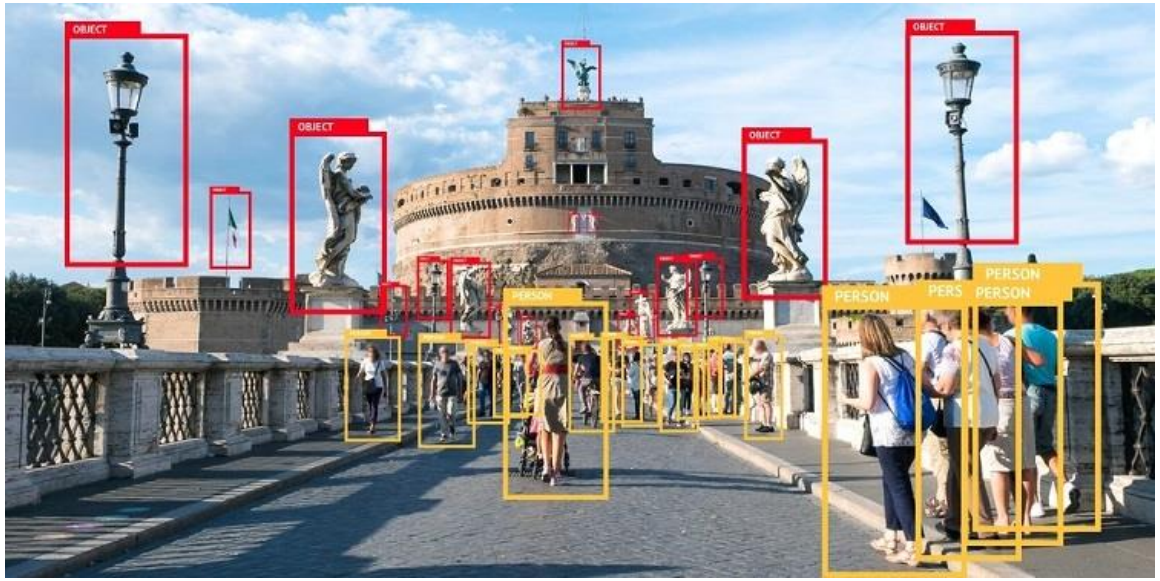
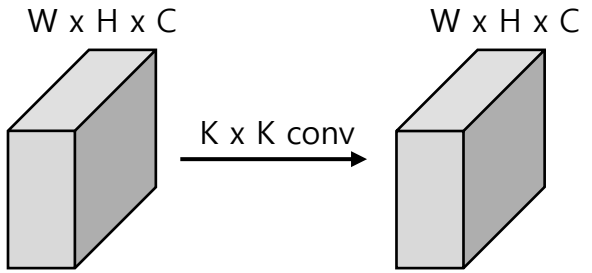
Swin Transformer Hierarchical Vision Transformer using Shifted Windows

Written in 2022.08.28

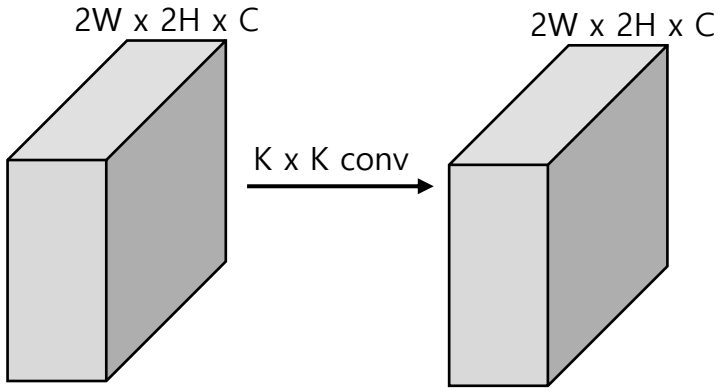
Research Background

- Vision Task에서는 객체들의 크기가 다르고, 입력 이미지가 고해상도일수록 연산량이 매우 늘어난다는 단점이 있음
- 이는 Vision Task에서 또한 발생함
- 다양한 크기의 객체를 위해 작은 Patch부터 계산을 진행하며, 점점 병합을 통해 큰 Patch를 확인하는 계층적 방식과 연산량이 늘어나지 않고 window의 상관관계를 구하는 SW-MSA를 제안함





연산량: (W x H x C) x (K x K x C) $=K^2WHC^2$



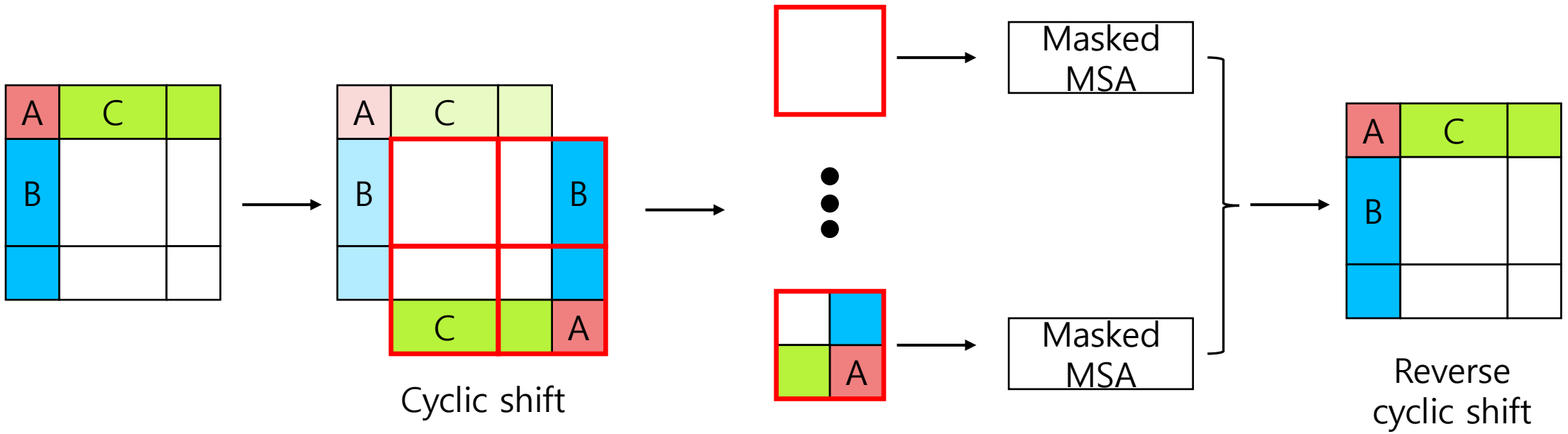
연산량: (W x H x C) x (2K x 2K x C) $=4K^2WHC^2$

Model

- 계층적 구조를 위해 크기가 작은 패치를 만들고 window내에서 Self attention을 진행한 후 주변의 window와 합쳐 Self attention 하는 것을 반복한다.

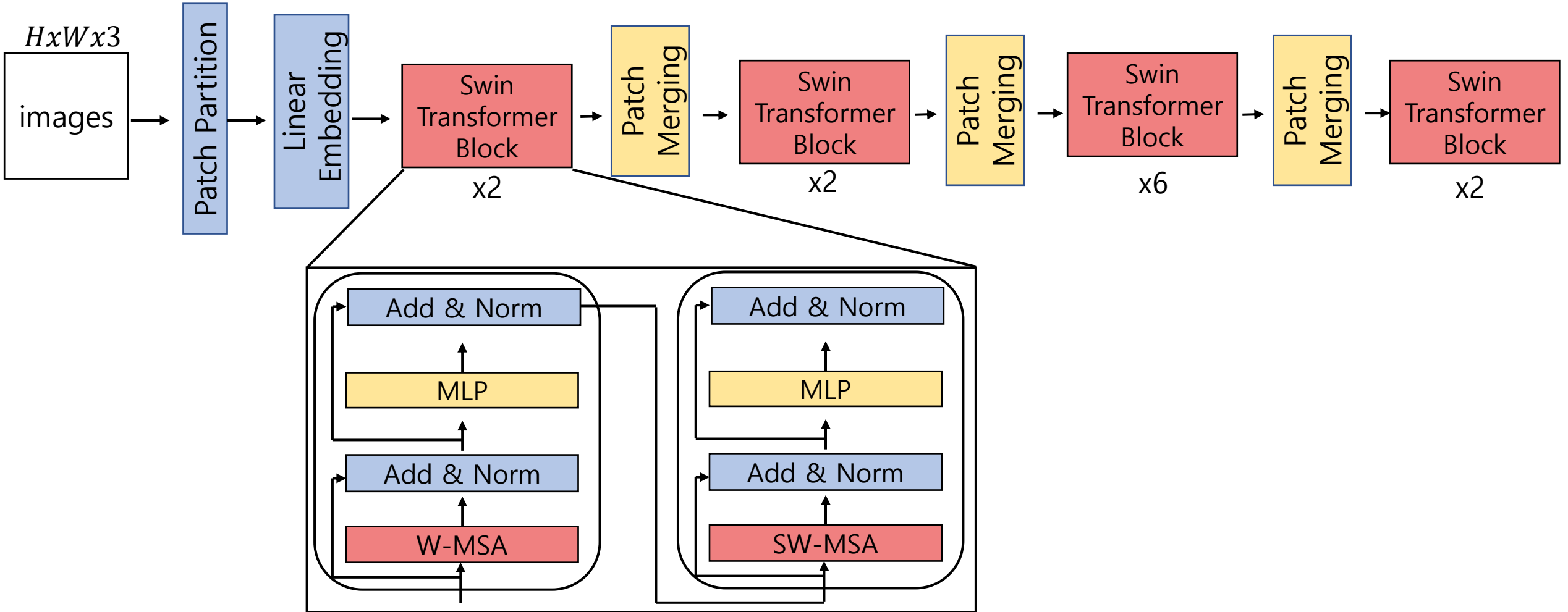
Model

- Window에 partition을 나누고, 각 window간의 상관 관계를 구하는 것은 Computer vision task에 도움이 됨
- 이를 위해 window를 shift하고 pad를 추가하면 연산량이 늘어남
- Pad하지 않고 window간의 상관 관계를 구할 수 있는 SW-MSA를 제안함
- Cyclic shift로 같은 window를 추가하고, 기존에 존재하는 window는 계산하지 않기 위해 masked MSA를 진행 후 reverse cyclic shift로 원래대로 돌림



Model

- 전체 모델 구조는 이미지가 들어가면 이를 patch로 나누고, Linear Embedding함
- 그리고 Swin Transformer Block과 Patch Merging을 하여 작은 패치부터 큰 패치까지 attention함



Result

- ImageNet Classification 에서 SOTA 달성함
- 그리고 ImageNet을 통해 학습된 모델을 Backbone으로 사용하였을 때, Object Detection, Segmentation 쪽에서도 SOTA를 달성함
- Vision Transformer와 같이 학습 데이터가 많아질 수록 Accuracy와 mAP가 증가하는 추세를 보임

(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

(b) ImageNet-22K pre-trained models

method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
R-101x3 [38]	384^2	388M	204.6G	-	84.4
R-152x4 [38]	480^2	937M	840.5G	-	85.4
ViT-B/16 [20]	384^2	86M	55.4G	85.9	84.0
ViT-L/16 [20]	384^2	307M	190.7G	27.3	85.2
Swin-B	224^2	88M	15.4G	278.1	85.2
Swin-B	384^2	88M	47.0G	84.7	86.4
Swin-L	384^2	197M	103.9G	42.1	87.3

한줄평

- YOLO나 SSD에서 다른 크기의 grid에서 예측을 진행해 큰 객체부터 작은 객체까지 탐지해내는 것을 알고있었지만 ViT를 에서 이런 방식이 쓰이는 것을 생각하지 못함
- 이것을 통해 이전의 연구들에 대한 이해가 중요하다 생각됨