

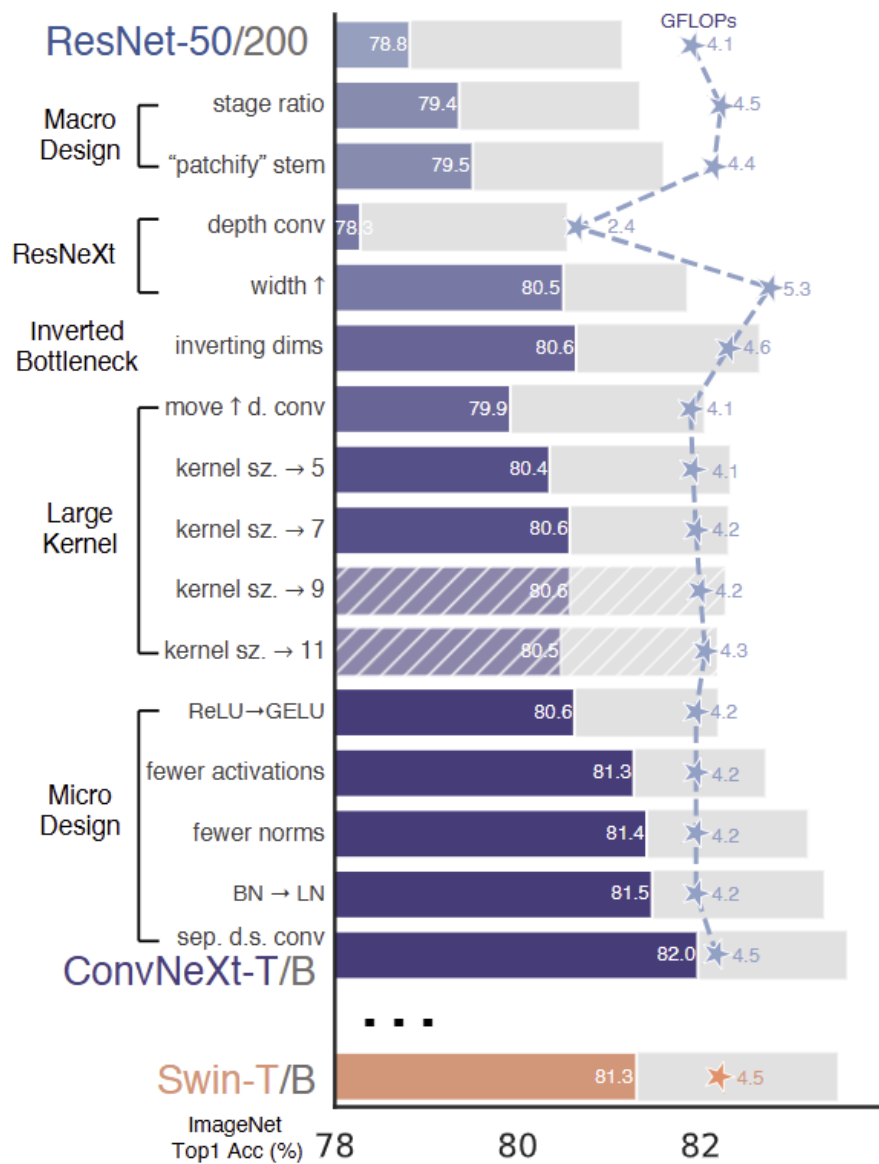
# A ConvNet for the 2020s

<https://arxiv.org/pdf/2201.03545>

## 0. Introduction

- Vision Transformer 계열이 이미지 분류에서 성능을 주도하며 CNN의 역할이 줄어들었다는 평가가 있었음.
- 하지만 Transformer 구조는 계산 비용과 메모리 요구량이 높고, 일부 비전 태스크에서는 여전히 한계가 존재함.
- 본 논문은 CNN 구조를 현대적으로 재설계하면 Transformer와 경쟁 가능한지 검증하고자 함.
- 이를 위해 기존 ResNet을 기반으로 구조를 개선한 **ConvNeXt** 모델을 제안함.

## 1. Overview



- ConvNeXt는 순수 Convolution 기반 구조로 최신 Transformer 모델과 경쟁 가능한 성능을 달성함.
- 큰 커널, depthwise convolution, LayerNorm 등 현대적 설계 요소를 CNN에 적용함.
- 이미지 분류뿐 아니라 detection, segmentation 등의 다운스트림 태스크에서도 높은 성능을 보임.

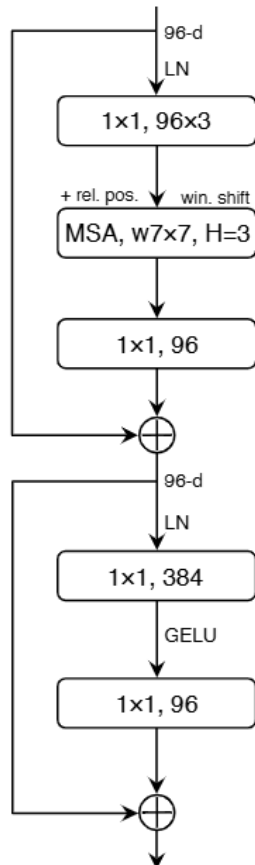
## 2. Challenges

- Transformer 구조는 모든 위치 간 관계 계산으로 인해 계산 비용이 큼.

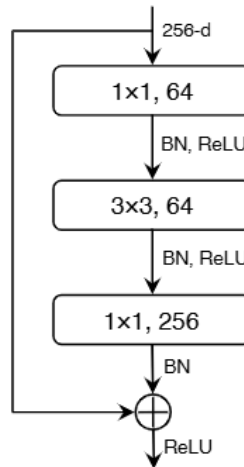
- 기존 CNN은 지역 정보에는 강하지만 전역 정보 표현에서 제한이 있었음.
- 논문은 Transformer 설계 요소 일부를 CNN에 통합해 이 문제를 완화하고자 함.

### 3. Method

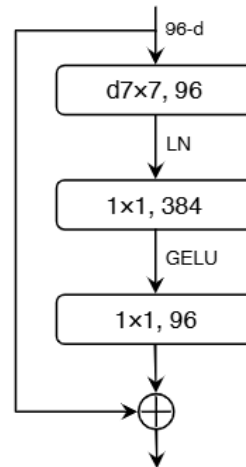
#### Swin Transformer Block



#### ResNet Block



#### ConvNeXt Block



#### Modernized ConvNet Design

- ResNet 구조를 기반으로 patchify stem, large kernel, inverted bottleneck 구조를 도입함.
- depthwise convolution을 사용하여 계산 효율을 유지하면서 표현력을 향상시킴.
- BatchNorm 대신 LayerNorm, 활성화 함수로 GELU를 적용함.

#### Architecture Adjustment

- 각 stage의 block 수와 채널 구조를 조정해 Transformer 모델과 유사한 계산량으로 구성함.

- CNN의 지역 특성은 유지하면서 더 넓은 receptive field 확보를 목표로 설계함.

## 4. Experiments

- ImageNet 분류 태스크에서 ConvNeXt 계열 모델의 성능을 평가함.
- COCO detection 및 ADE20K segmentation 등의 다운스트림 태스크에서도 실험 수행함.
- 다양한 모델 크기 설정에서 성능을 비교 분석함.

## 5. Results

model	image size	#param.	FLOPs	throughput (image / s)	IN-1K top-1 acc.
ImageNet-1K trained models					
● RegNetY-16G [54]	224 <sup>2</sup>	84M	16.0G	334.7	82.9
● EffNet-B7 [71]	600 <sup>2</sup>	66M	37.0G	55.1	84.3
● EffNetV2-L [72]	480 <sup>2</sup>	120M	53.0G	83.7	85.7
○ DeiT-S [73]	224 <sup>2</sup>	22M	4.6G	978.5	79.8
○ DeiT-B [73]	224 <sup>2</sup>	87M	17.6G	302.1	81.8
○ Swin-T	224 <sup>2</sup>	28M	4.5G	757.9	81.3
● ConvNeXt-T	224 <sup>2</sup>	29M	4.5G	774.7	<b>82.1</b>
○ Swin-S	224 <sup>2</sup>	50M	8.7G	436.7	83.0
● ConvNeXt-S	224 <sup>2</sup>	50M	8.7G	447.1	<b>83.1</b>
○ Swin-B	224 <sup>2</sup>	88M	15.4G	286.6	83.5
● ConvNeXt-B	224 <sup>2</sup>	89M	15.4G	292.1	<b>83.8</b>
○ Swin-B	384 <sup>2</sup>	88M	47.1G	85.1	84.5
● ConvNeXt-B	384 <sup>2</sup>	89M	45.0G	95.7	<b>85.1</b>
● ConvNeXt-L	224 <sup>2</sup>	198M	34.4G	146.8	<b>84.3</b>
● ConvNeXt-L	384 <sup>2</sup>	198M	101.0G	50.4	<b>85.5</b>
ImageNet-22K pre-trained models					
● R-101x3 [39]	384 <sup>2</sup>	388M	204.6G	-	84.4
● R-152x4 [39]	480 <sup>2</sup>	937M	840.5G	-	85.4
● EffNetV2-L [72]	480 <sup>2</sup>	120M	53.0G	83.7	86.8
● EffNetV2-XL [72]	480 <sup>2</sup>	208M	94.0G	56.5	87.3
○ ViT-B/16 (🐼) [67]	384 <sup>2</sup>	87M	55.5G	93.1	85.4
○ ViT-L/16 (🐼) [67]	384 <sup>2</sup>	305M	191.1G	28.5	86.8
● ConvNeXt-T	224 <sup>2</sup>	29M	4.5G	774.7	<b>82.9</b>
● ConvNeXt-T	384 <sup>2</sup>	29M	13.1G	282.8	<b>84.1</b>
● ConvNeXt-S	224 <sup>2</sup>	50M	8.7G	447.1	<b>84.6</b>
● ConvNeXt-S	384 <sup>2</sup>	50M	25.5G	163.5	<b>85.8</b>
○ Swin-B	224 <sup>2</sup>	88M	15.4G	286.6	85.2
● ConvNeXt-B	224 <sup>2</sup>	89M	15.4G	292.1	<b>85.8</b>
○ Swin-B	384 <sup>2</sup>	88M	47.0G	85.1	86.4
● ConvNeXt-B	384 <sup>2</sup>	89M	45.1G	95.7	<b>86.8</b>
○ Swin-L	224 <sup>2</sup>	197M	34.5G	145.0	86.3
● ConvNeXt-L	224 <sup>2</sup>	198M	34.4G	146.8	<b>86.6</b>
○ Swin-L	384 <sup>2</sup>	197M	103.9G	46.0	87.3
● ConvNeXt-L	384 <sup>2</sup>	198M	101.0G	50.4	<b>87.5</b>
● ConvNeXt-XL	224 <sup>2</sup>	350M	60.9G	89.3	<b>87.0</b>
● ConvNeXt-XL	384 <sup>2</sup>	350M	179.0G	30.2	<b>87.8</b>

backbone	FLOPs	FPS	AP <sup>box</sup>	AP <sup>box</sup> <sub>50</sub>	AP <sup>box</sup> <sub>75</sub>	AP <sup>mask</sup>	AP <sup>mask</sup> <sub>50</sub>	AP <sup>mask</sup> <sub>75</sub>
Mask-RCNN 3× schedule								
○ Swin-T	267G	23.1	46.0	68.1	50.3	41.6	65.1	44.9
● ConvNeXt-T	262G	25.6	<b>46.2</b>	67.9	50.8	<b>41.7</b>	65.0	44.9
Cascade Mask-RCNN 3× schedule								
● ResNet-50	739G	16.2	46.3	64.3	50.5	40.1	61.7	43.4
● X101-32	819G	13.8	48.1	66.5	52.4	41.6	63.9	45.2
● X101-64	972G	12.6	48.3	66.4	52.3	41.7	64.0	45.1
○ Swin-T	745G	12.2	50.4	69.2	54.7	43.7	66.6	47.3
● ConvNeXt-T	741G	13.5	<b>50.4</b>	69.1	54.8	<b>43.7</b>	66.5	47.3
○ Swin-S	838G	11.4	51.9	70.7	56.3	45.0	68.2	48.8
● ConvNeXt-S	827G	12.0	<b>51.9</b>	70.8	56.5	<b>45.0</b>	68.4	49.1
○ Swin-B	982G	10.7	51.9	70.5	56.4	45.0	68.1	48.9
● ConvNeXt-B	964G	11.4	<b>52.7</b>	71.3	57.2	<b>45.6</b>	68.9	49.5
○ Swin-B <sup>‡</sup>	982G	10.7	53.0	71.8	57.5	45.8	69.4	49.7
● ConvNeXt-B <sup>‡</sup>	964G	11.5	<b>54.0</b>	73.1	58.8	<b>46.9</b>	70.6	51.3
○ Swin-L <sup>‡</sup>	1382G	9.2	53.9	72.4	58.8	46.7	70.1	50.8
● ConvNeXt-L <sup>‡</sup>	1354G	10.0	<b>54.8</b>	73.8	59.8	<b>47.6</b>	71.3	51.7
● ConvNeXt-XL <sup>‡</sup>	1898G	8.6	<b>55.2</b>	74.2	59.9	<b>47.7</b>	71.6	52.2

backbone	input crop.	mIoU	#param.	FLOPs
ImageNet-1K pre-trained				
○ Swin-T	512 <sup>2</sup>	45.8	60M	945G
● ConvNeXt-T	512 <sup>2</sup>	<b>46.7</b>	60M	939G
○ Swin-S	512 <sup>2</sup>	49.5	81M	1038G
● ConvNeXt-S	512 <sup>2</sup>	<b>49.6</b>	82M	1027G
○ Swin-B	512 <sup>2</sup>	49.7	121M	1188G
● ConvNeXt-B	512 <sup>2</sup>	<b>49.9</b>	122M	1170G
ImageNet-22K pre-trained				
○ Swin-B <sup>‡</sup>	640 <sup>2</sup>	51.7	121M	1841G
● ConvNeXt-B <sup>‡</sup>	640 <sup>2</sup>	<b>53.1</b>	122M	1828G
○ Swin-L <sup>‡</sup>	640 <sup>2</sup>	53.5	234M	2468G
● ConvNeXt-L <sup>‡</sup>	640 <sup>2</sup>	<b>53.7</b>	235M	2458G
● ConvNeXt-XL <sup>‡</sup>	640 <sup>2</sup>	<b>54.0</b>	391M	3335G

- ConvNeXt는 Transformer 기반 모델과 유사하거나 더 높은 성능을 달성함.
- 계산 효율과 구조 단순성을 유지하면서도 높은 정확도를 확보함.
- 다양한 비전 태스크에서 안정적인 성능 향상을 확인함.

## 6. Insight

- CNN 구조는 여전히 현대 비전 모델에서 경쟁력이 있음을 보여줌.
- Transformer 설계 요소 일부를 CNN에 통합하는 접근이 효과적임을 확인함.
- 구조가 단순하고 효율적이므로 실무 적용성이 높음.
- 향후 self-supervised 학습이나 다른 학습 방식과 결합 시 추가 발전 가능성이 존재함.