2023 하계 워크샵

경영과학연구실 김윤석

Index

Introduction

- CNN은 Computer vision 영역에서 필수적인 신경망으로 취급되어 왔음
- ViT의 성공은 다양한 시각에서의 연구를 촉발시킴
- ViT의 성공이 self-attentio에 의한 것인지, 패치 입력에 의한것인지, 대규모 데이터 학습에 의한 것인지에 대한 연구가 진행되기 시작함
- Computer vision 영역에서 Transformer는 다양한 변형이 제안되며 활발히 연구되기 시작함

MLP-Mixer: An all-MLP Architecture for vision

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer

Google Research

Problem statement

- 단순하고 효율적 연산이 가능한 Computer vision 모델에 대해 연구함

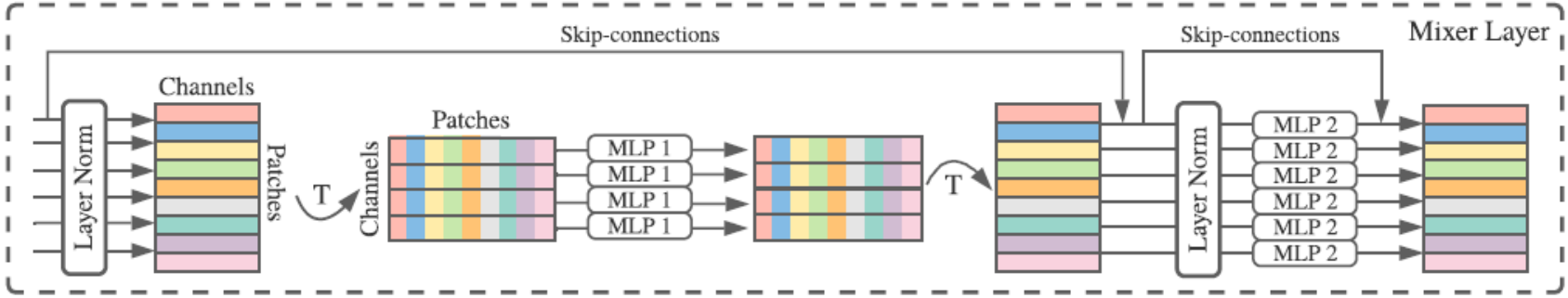
- Issue

  ViT는 patch에 대한 global attention 연산 수행으로 연산 비용이 큼

  CNN은 지속적 발전으로 크고 복잡한 구조를 보임

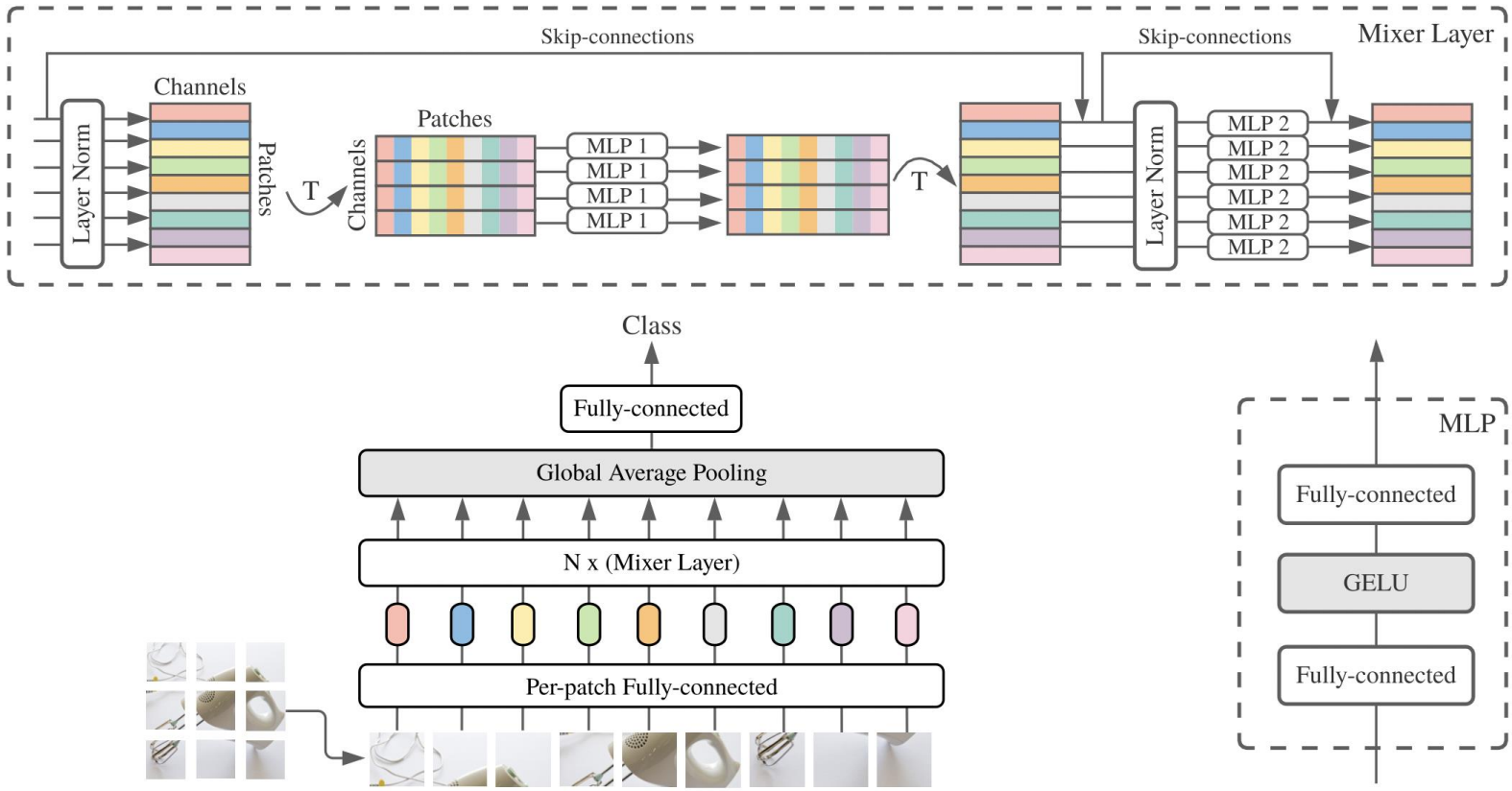
Key idea

- 두 가지 MLP layer의 입력을 다르게 함으로 간단한 MLP 구조로 이미지의 특징을 학습 할 수 있는 모델을 제안함



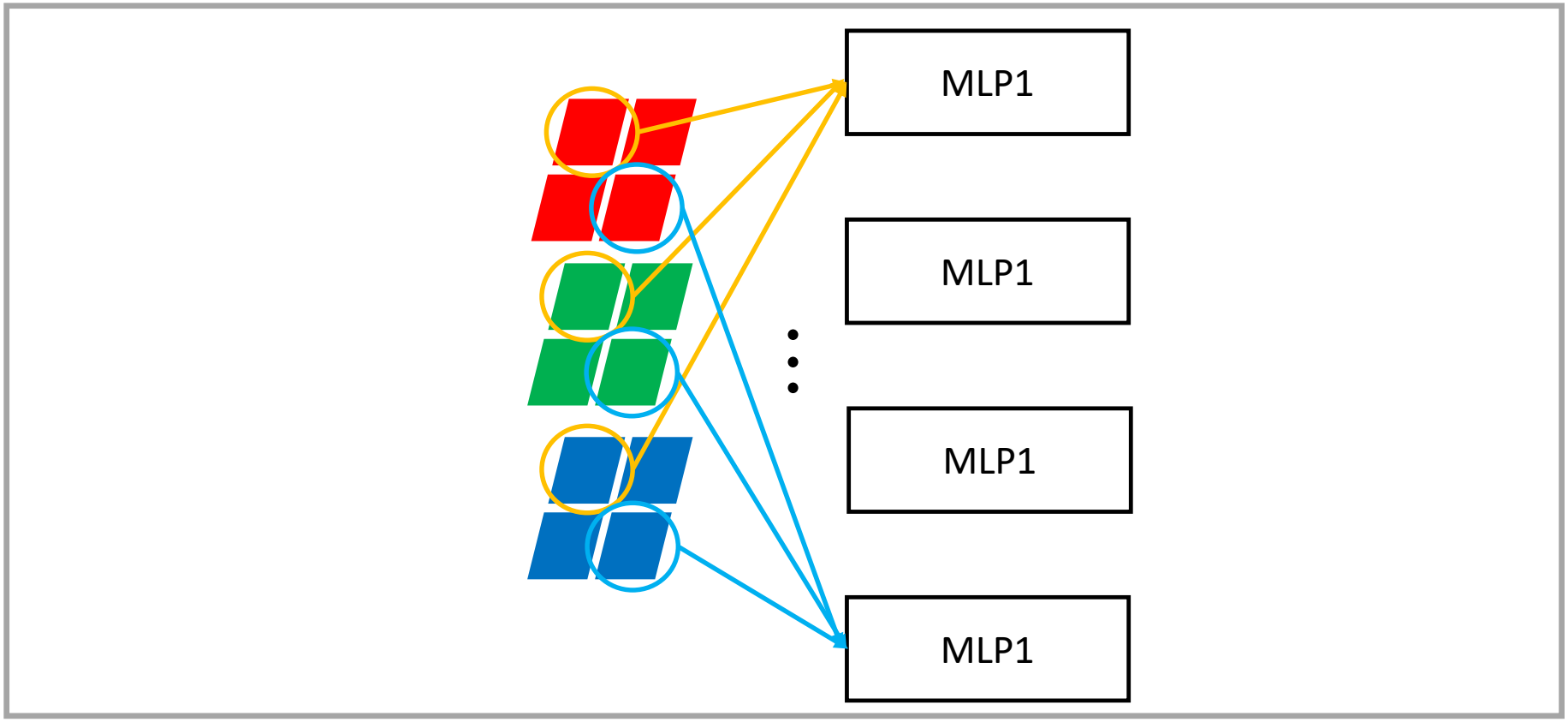
Model Architecture

- MLP-Mixer는 token-mixing mlp와 channel-mixing mlp 두개로 이루어져 있음



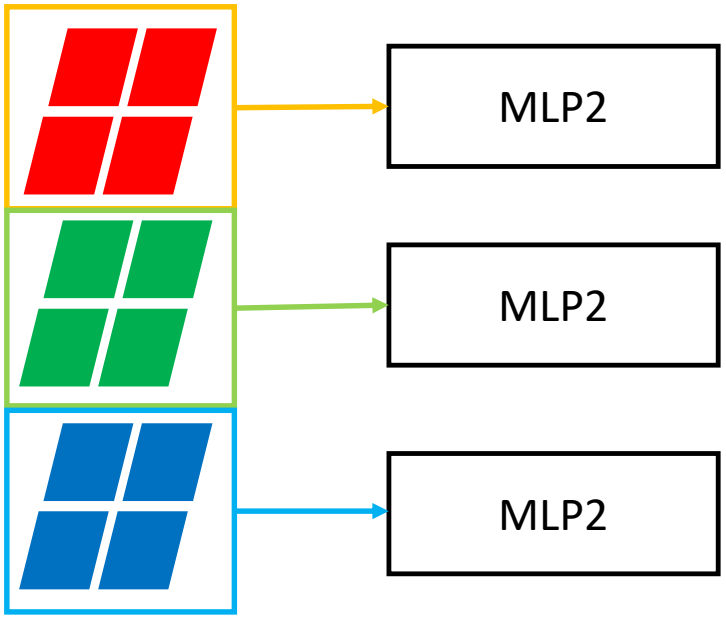
Channel-mixing MLP

- Channel-mixing MLP는 같은 공간의 특징을 혼합하는 MLP layer임
- ConvNet에서 필터를 통해 채널간 연산을 하는 과정과 같음



Token-mixing MLP

- Token-mixing MLP는 서로 다른 공간의 특징을 혼합하는 MLP layer임
- ConvNet에서 필터를 통해 픽셀을 줄이는 과정과 같음



Experiment result

- MLP-Mixer는 CNN model과 Vision Transformer와 비교하는 실험을 진행함
- BiT-R#xC: 이 모델은 Big Transfer를 적용한 ResNet#이며 큰 학습을 위해 채널을 C배 확장한 모델임.

	Image size	Pre-Train Epochs	ImNet top-1	ReaL top-1	Avg. 5 top-1	Throughput (img/sec/core)	TPUv3 core-days
Pre-trained on ImageNet (with extra regularization)							
Mixer-B/16	224	300	76.44	82.36	88.33	1384	0.01k(‡)
ViT-B/16 (☎)	224	300	79.67	84.97	90.79	861	0.02k(‡)
Mixer-L/16	224	300	71.76	77.08	87.25	419	0.04k(‡)
ViT-L/16 (☎)	224	300	76.11	80.93	89.66	280	0.05k(‡)
Pre-trained on ImageNet-21k (with extra regularization)							
Mixer-B/16	224	300	80.64	85.80	92.50	1384	0.15k(‡)
ViT-B/16 (☎)	224	300	84.59	88.93	94.16	861	0.18k(‡)
Mixer-L/16	224	300	82.89	87.54	93.63	419	0.41k(‡)
ViT-L/16 (☎)	224	300	84.46	88.35	94.49	280	0.55k(‡)
Mixer-L/16	448	300	83.91	87.75	93.86	105	0.41k(‡)
Pre-trained on JFT-300M							
Mixer-S/32	224	5	68.70	75.83	87.13	11489	0.01k
Mixer-B/32	224	7	75.53	81.94	90.99	4208	0.05k
Mixer-S/16	224	5	73.83	80.60	89.50	3994	0.03k
BiT-R50x1	224	7	73.69	81.92	—	2159	0.08k
Mixer-B/16	224	7	80.00	85.56	92.60	1384	0.08k
Mixer-L/32	224	7	80.67	85.62	93.24	1314	0.12k
BiT-R152x1	224	7	79.12	86.12	—	932	0.14k
BiT-R50x2	224	7	78.92	86.06	—	890	0.14k
BiT-R152x2	224	14	83.34	88.90	—	356	0.58k
Mixer-L/16	224	7	84.05	88.14	94.51	419	0.23k
Mixer-L/16	224	14	84.82	88.48	94.77	419	0.45k
ViT-L/16	224	14	85.63	89.16	95.21	280	0.65k
Mixer-H/14	224	14	86.32	89.14	95.49	194	1.01k
BiT-R200x3	224	14	84.73	89.58	—	141	1.78k
Mixer-L/16	448	14	86.78	89.72	95.13	105	0.45k
ViT-H/14	224	14	86.65	89.56	95.57	87	2.30k
ViT-L/16 [14]	512	14	87.76	90.54	95.63	32	0.65k

Conclusion

- MLP-Mixer는 최근 대규모 데이터셋을 활용한 전이학습에서 CNN과 ViT와 경쟁력이 있음
- MLP 구조만 활용하는 점에서 CNN 또는 ViT보다 간단하며 효율적일 수 있음
- 대규모 데이터셋을 활용한 전이학습에 대해서만 실험을 진행하여 MLP-Mixer가 실제로 활용되기 어려워 보임

PATCHES ARE ALL YOU NEED? 🤷

Asher Trockman, J. Zico Kolter

Carnegie Mellon University and Bosch Center for AI

Problem statement

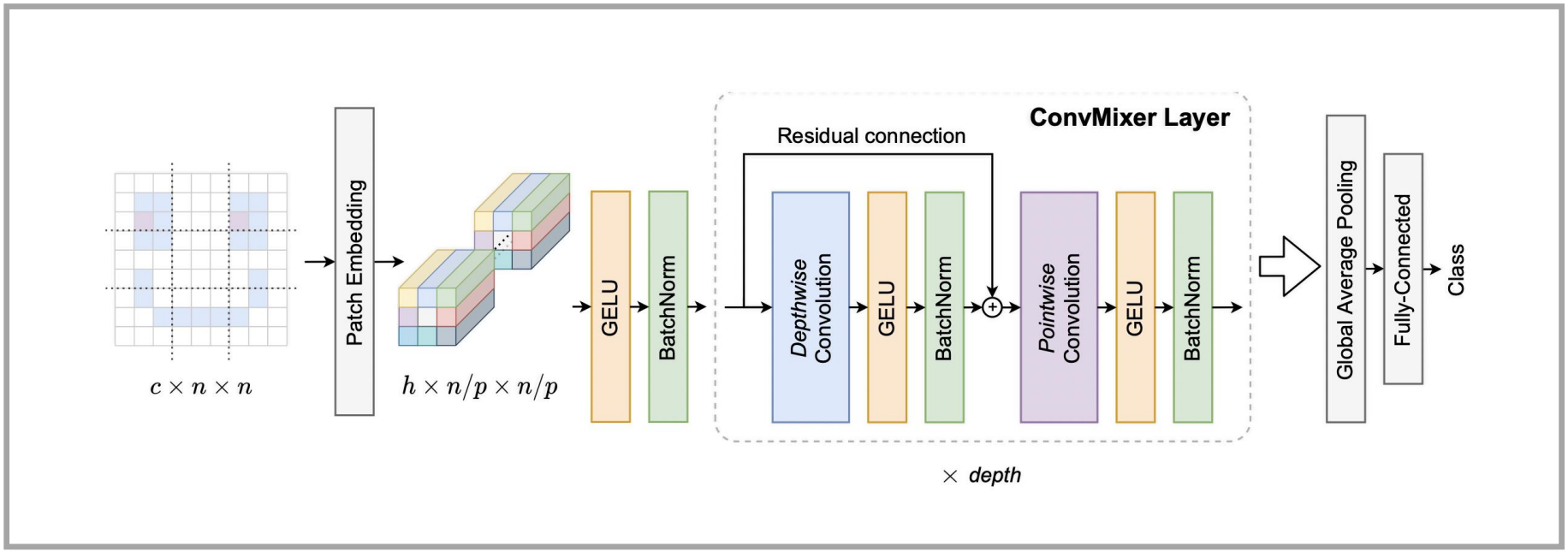
- ViT의 성공이 self-attention 매커니즘 보다 패치 기반 표현이 중요함을 보여야 함
- MLP-Mixer와 마찬가지로 transformer layer는 사용하지 않고 패치 기반 표현의 효과를 분석해야함

Key ideas

- Convolution을 이용한 패치 임베딩 방법을 제안함
- Depthwise convolution과 Pointwise convolution을 활용하여 간단한 패치 기반 학습 모델을 제안함

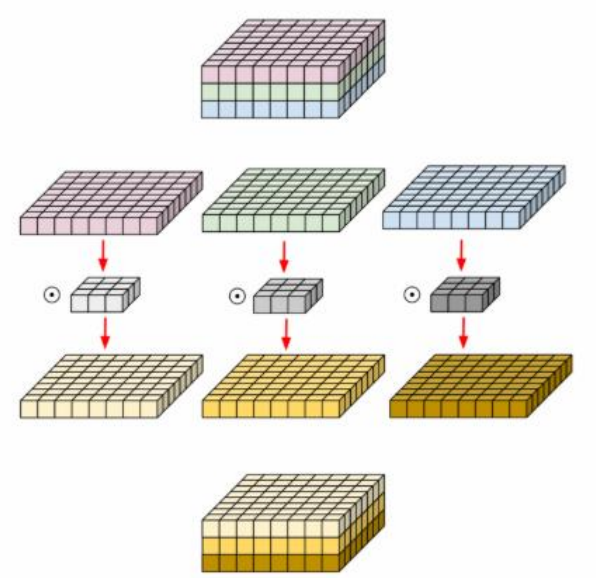
Model architecture

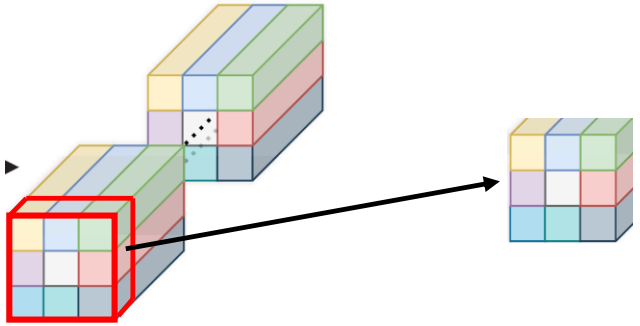
- ConvMixer는 Embedding layer, ConvMixer layer, Classification layer로 구성되어 있음
- ConvMixer layer는 Depthwise Convolution과 Pointwise Convolution으로 구성되어 있음



Depthwise Convolution

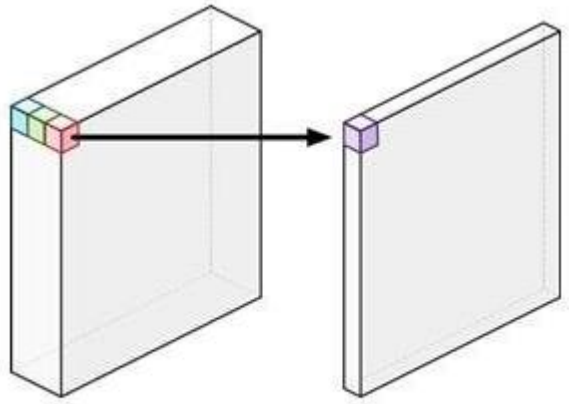
- Depthwise Convolution은 채널 별로 convolution 연산을 수행함
- Depthwise Convolution을 통해 나온 특징 맵은 채널 수가 유지됨
- Depthwise Convolution은 공간적 특징을 학습하게 가능하게 함



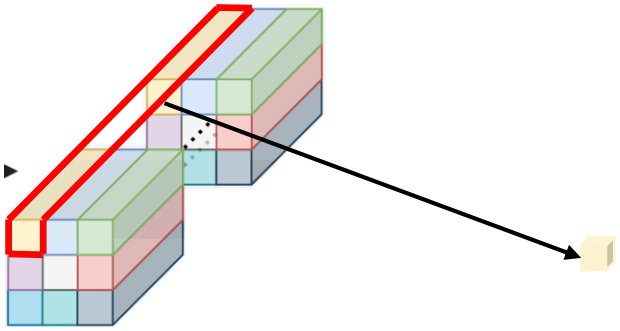


Pointwise Convolution

- Pointwise Convolution은 1x1 크기의 필터를 사용하는 convolution layer임
- Pointwise Convolution은 입력의 모든 채널 정보를 통합하는 역할을 함



b) pointwise convolution



Experiment result

- ConvMixer는 DeiT, ResNet, ResMLP와 비교하는 실험을 진행함
- DeiT는 Kknowledge distilation을 활용한 ViT로 대규모 데이터 학습 없이 경쟁력 있는 성능을 보임
- ResMLP는 MLP layer만 사용하여 이미지를 처리하는 모델임

Current "Most Interesting" **ConvMixer** Configurations *vs.* Other Simple Models

Network	Patch Size	Kernel Size	# Params ($\times 10^6$)	Throughput (img/sec)	Act. Fn.	# Epochs	ImNet top-1 (%)
ConvMixer-1536/20	7	9	51.6	134	G	150	81.37
ConvMixer-768/32	7	7	21.1	206	R	300	80.16
ResNet-152	–	3	60.2	828	R	150	79.64
DeiT-B	16	–	86	792	G	300	81.8
ResMLP-B24/8	8	–	129	181	G	400	81.0

Conclusion

- ConvMixer는 ViT, MLP-Mixer와 다르게 대규모 데이터 학습 없이 경쟁력 있는 성능을 보임
- ConvMixer는 ViT의 성공적 등장을 Self-attention 매커니즘이 아닌 patch embedding 또는 대규모 데이터 학습일 수 있음을 제시함
- ConvMixer는 전통적인 신경망 설계를 따르지 않고도 효과적인 비전 모델을 설계 할 수 있음을 보여주는 연구임

DaViT: Dual Attention Vision Transformers

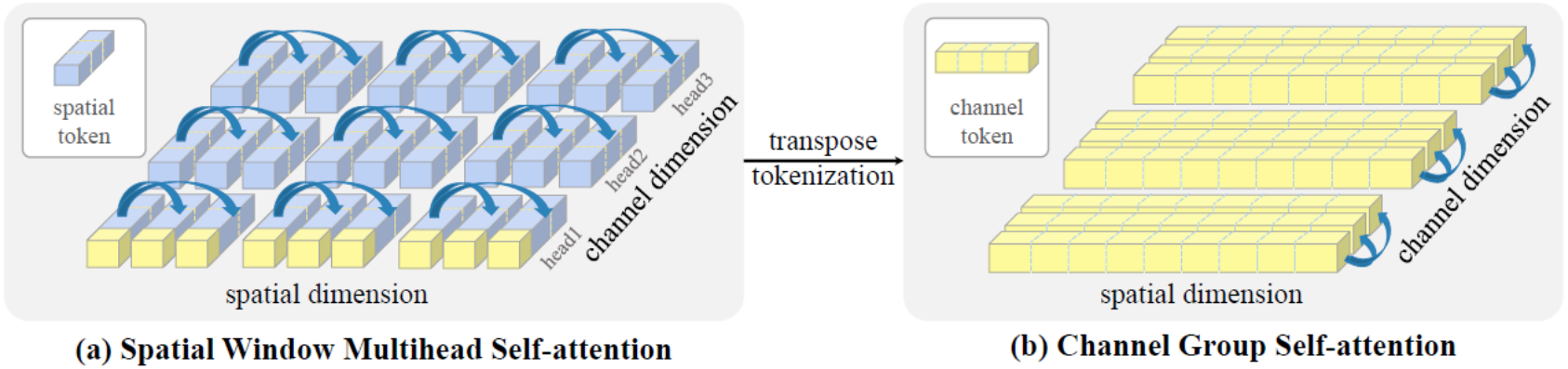
Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, Lu Yuan
The University of Hong Kong, Microsoft Cloud + AI

Problem Statement

- ViT는 계산 비용과 입력 공간에 따른 트레이드 오프가 존재함
- DaViT은 효율적인 입력 공간과 연산 비용을 갖는 Vision Transformer를 연구함

Key idea

- 패치간 정보를 학습하는 Spatial Window Self attentio과 패치에서 채널간 정보를 학습하는 Channel Group Self attention Dual Attention Block 제안함

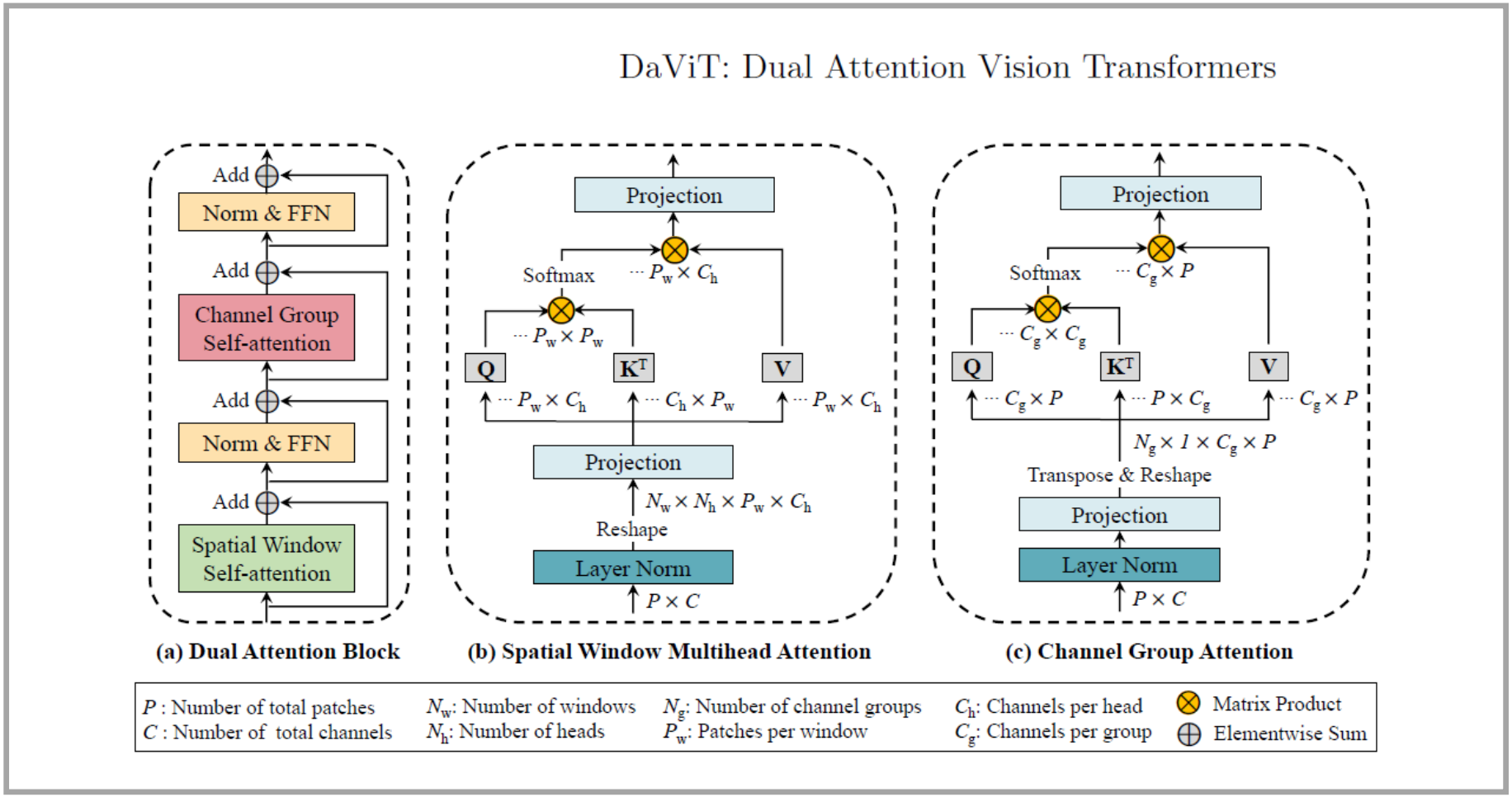


(a) Spatial Window Multihead Self-attention

(b) Channel Group Self-attention

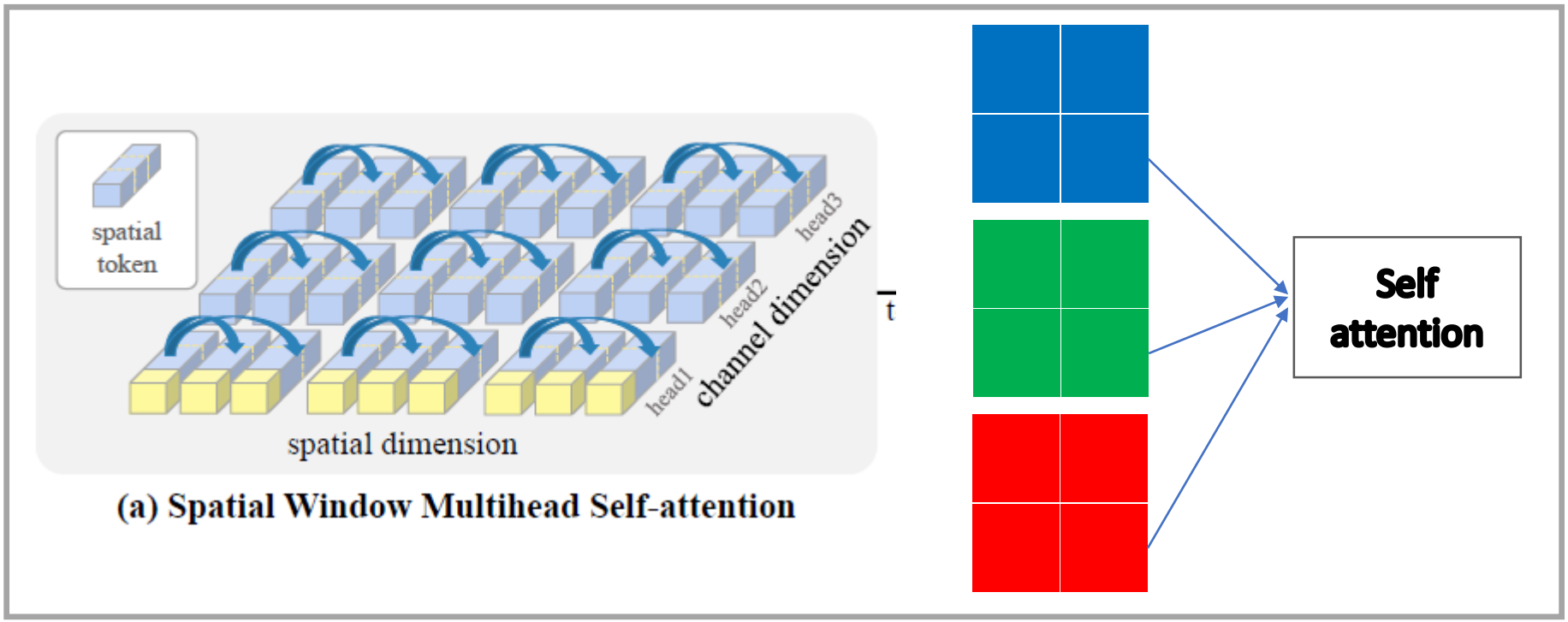
Model architecture

- DaViT은 Spatial Window Self-attention layer와 Channel Group Self-attention으로 이루어져 있음



Spatial Window Self-attention

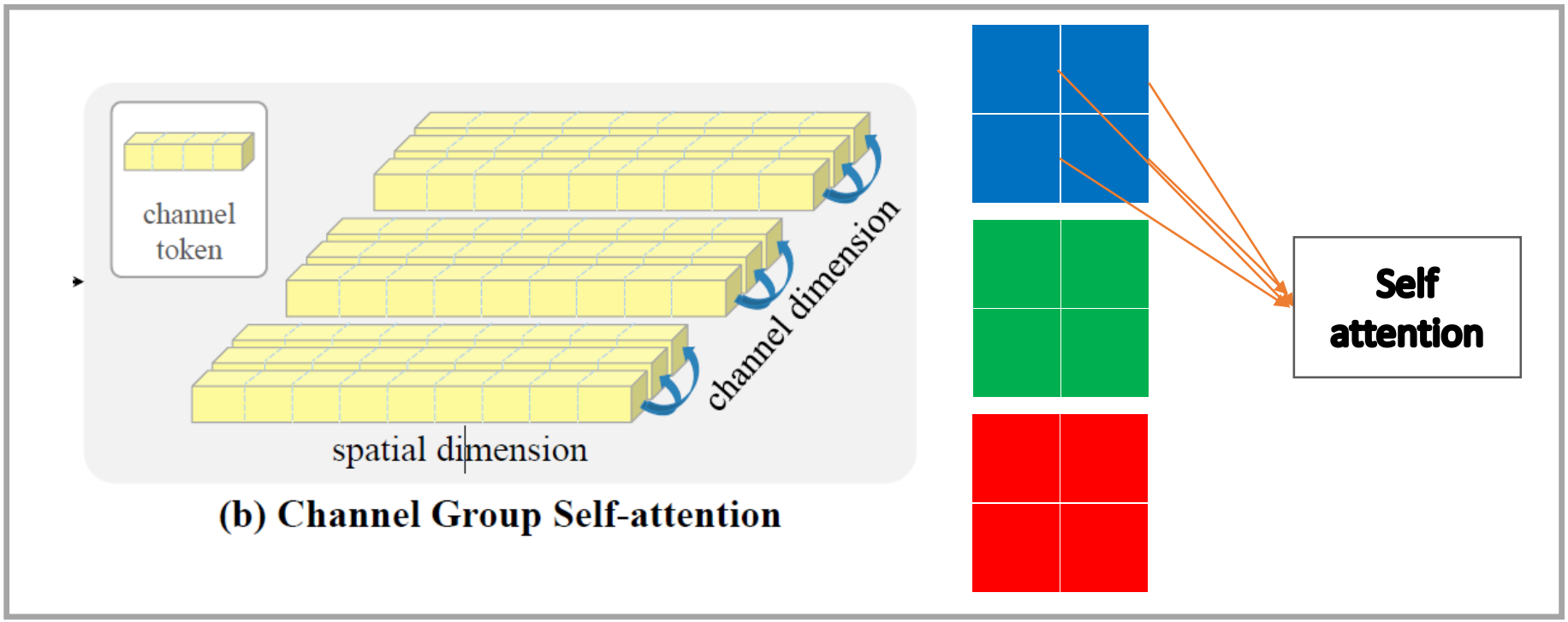
- Spatial window self-attention은 윈도우 내의 패치간의 Self-attention 연산을 수행하는 layer임
- Spatial window self-attention은 지역 특징을 학습하는 역할을 하지만 전역 특징을 학습하지 못함



(a) Spatial Window Multihead Self-attention

Channel Group Self-attention

- Channel group self-attention은 같은 채널에서 self-attention을 수행하는 layer임
- Channel group self-attention은 전역 정보를 학습하는 역할을 함



Experiment result

- DaViT는 다양한 Transformer 모델과 비교하는 실험을 진행함
- DaViT는 비슷한 규모의 모델에서 가장 높은 성능을 보이며 복잡도 또한 낮음

Model	#Params (M)	FLOPs (G)	Top-1 (%)
ResNet-50 [24]	25.0	4.1	76.2
DeiT-Small/16 [55]	22.1	4.5	79.8
PVT-Small [63]	24.5	3.8	79.8
ConvMixer-768/32 [57]	21.1	–	80.2
CrossViT-Small [5]	26.7	5.6	81.0
Swin-Tiny [40]	28.3	4.5	81.2
CvT-13 [65]	20.0	4.5	81.6
CoAtNet-0 [13]	25.0	4.2	81.6
CaiT-XS-24 [56]	26.6	5.4	81.8
ViL-Small [82]	24.6	5.1	82.0
PVTv2-B2 [62]	25.4	4.0	82.0
UFO-ViT-S [50]	21.0	3.7	82.0
Focal-Tiny [72]	29.1	4.9	82.2
DaViT-Tiny (Ours)	28.3	4.5	**82.8**
ResNet-101 [24]	45.0	7.9	77.4
PVT-Medium [63]	44.2	6.7	81.2
CvT-21 [65]	32.0	7.1	82.5
UFO-ViT-M [50]	37.0	7.0	82.8
Swin-Small [40]	49.6	8.7	83.1
ViL-Medium [82]	39.7	9.1	83.3
CaiT-S36 [56]	68.0	13.9	83.3
CoAtNet-1 [13]	42.0	8.4	83.3
Focal-Small [72]	51.1	9.1	83.5
CSwin-S [17]	35.0	6.9	83.6
VAN-Large [21]	44.8	9.0	83.9
UniFormer-B [32]	50.0	8.3	83.9
DaViT-Small (Ours)	49.7	8.8	**84.2**

Model	#Params (M)	FLOPs (G)	Top-1 (%)
ResNet-152 [24]	60.0	11.0	78.3
PVT-Large [63]	61.4	9.8	81.7
DeiT-Base/16 [55]	86.7	17.4	81.8
CrossViT-Base [5]	104.7	21.2	82.2
T2T-ViT-24 [75]	64.1	14.1	82.3
CPVT-Base [11]	88.0	17.6	82.3
TNT-Base [22]	65.6	14.1	82.8
ViL-Base [82]	55.7	13.4	83.2
UFO-ViT-B [50]	64.0	11.9	83.3
Swin-Base [40]	87.8	15.4	83.4
CaiT-M24 [56]	185.9	36.0	83.4
NFNet-F0 [4]	71.5	12.4	83.6
PVTv2-B5 [62]	82.0	11.8	83.8
Focal-Base [72]	89.8	16.0	83.8
CoAtNet-2 [13]	75.0	15.7	84.1
CSwin-B [17]	78.0	15.0	84.2
DaViT-Base (Ours)	87.9	15.5	**84.6**
Pre-trained on ImageNet-22k			
Swin-Large [40] †	197.0	103.9	86.4
CSWin-B [17] †	78.0	47.0	87.0
CSWin-L [17] †	173.0	96.8	87.5
CoAtNet-3 [13] †	168.0	107.4	87.6
DaViT-Base (Ours) †	87.9	46.4	86.9
DaViT-Large (Ours) †	196.8	103.0	87.5
Pre-trained on 1.5B image and text pairs			
DaViT-Huge (Ours) ‡	362	334	90.2
DaViT-Giant (Ours) ‡	1437	1038	**90.4**

Conclusion

- 저자는 이미지의 전역 정보와 지역 정보를 함께 학습 가능하며 비용을 줄인 DaViT을 제안함
- DaViT은 spatial window self-attention과 channel group self-attention을 이용하여 전역 및 지역 정보를 학습하며 계산적으로 효율적인 layer를 구성함
- 서로 다른 attention layer를 통해 vision task에서 attention machnism에 새로운 방향을 제시함

QnA

Thank you