MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer

패턴인식

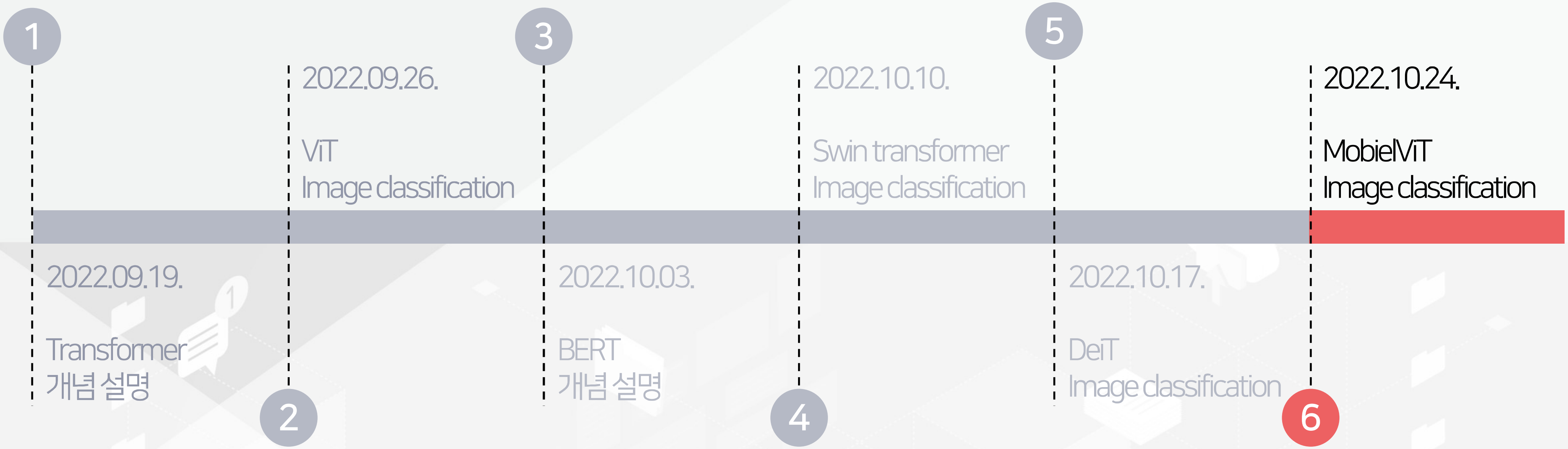
202132032 김형범

2022. 09. 19.

Mehta, Sachin, and Mohammad Rastegari. "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer." proc. of. ICLR 2022

CONTENTS

Transformer 주차별 계획



1

2022.09.19.

Transformer
개념 설명

2

2022.09.26.

ViT
Image classification

3

2022.10.03.

BERT
개념 설명

4

2022.10.10.

Swin transformer
Image classification

5

2022.10.17.

DeiT
Image classification

6

2022.10.24.

MobielViT
Image classification

BERT

CONTENTS

MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer

I Introduction + contributions

2 Method

3 Experiments

IV

Transformer

1 _ Introduction + contributions

1. ViT to Mobile-ViT

CHAPTER

I _

Introduction + contributions

ViT to Mobile-ViT

https://github.com/lucidrains/vit-pytorch

Introduction + contributions

ViT to Mobile-ViT

Mobile-ViT

Mobile과 같은 하드웨어의 자원이 제한된 곳에서 범용적으로(General purpose) 사용할 수 있도록 만든 작고(Light-weight) 빠른(Low-latency) ViT 모델

Introduction + contributions

_ ViT to Mobile-ViT

CNN

- Spatial(local) inductive bias

- data augmentation에 덜 민감

ViT

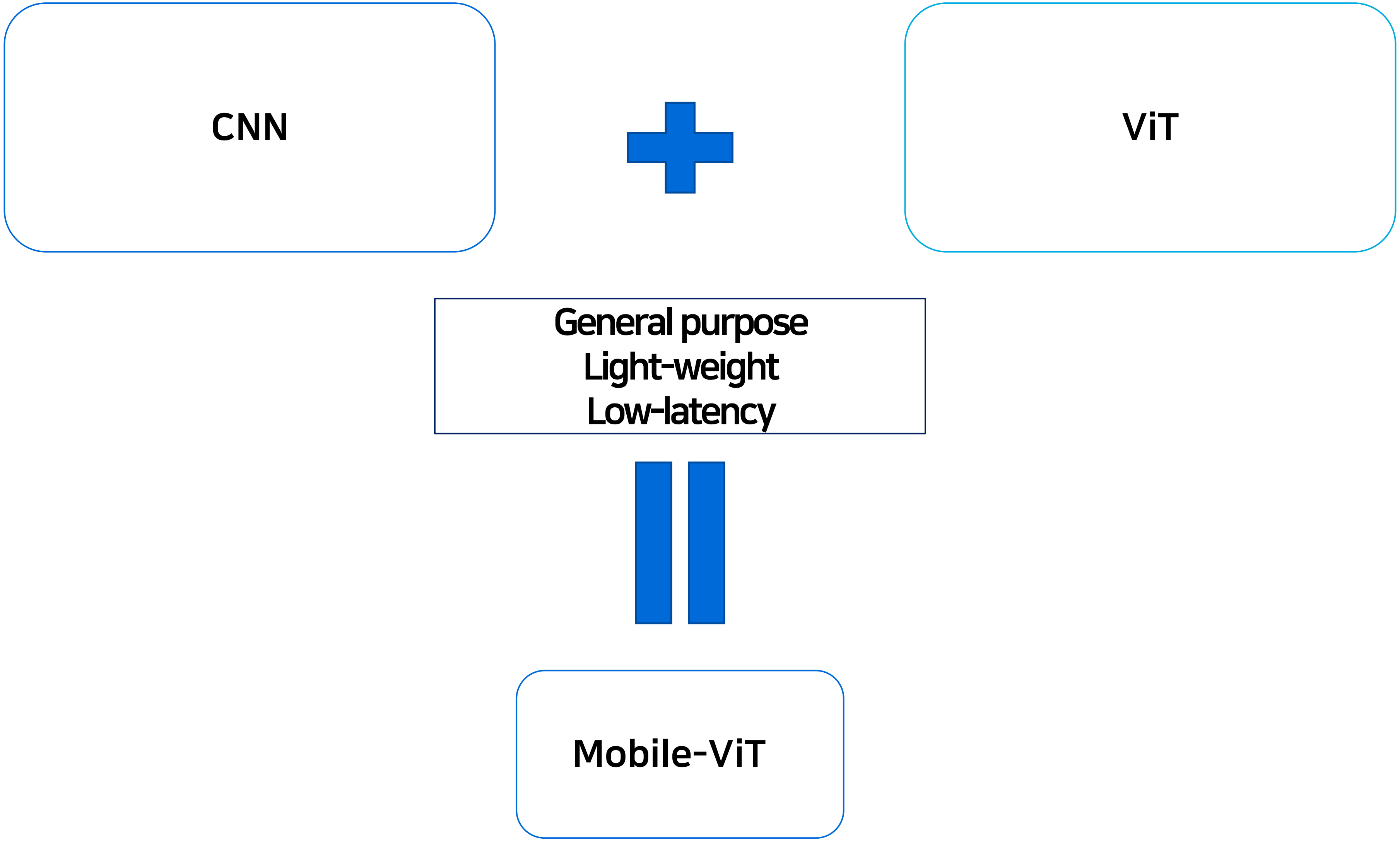
- Input-adaptive weighting

- global processing

CHAPTER I _

Introduction + contributions

ViT to Mobile-ViT



CNN

ViT

General purpose
Light-weight
Low-latency

Mobile-ViT

Introduction + contributions
ViT to Mobile-ViT

Generalization Problem

- Models are Brittle : 아무리 같은 의미의 데이터라도 조금만 바뀌면 모델이 망가진다.

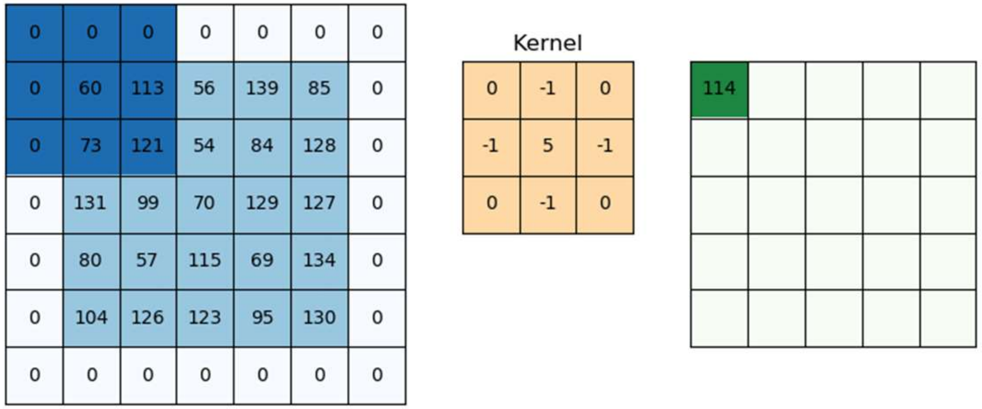
- Models are Spurious : 데이터의 진정한 의미를 파악하지 못하고 결과(Arifacts)와 편향(Bias)를 암기한다.

Inductive Bias

학습 시에는 만나보지 않았던 상황에 대하여 정확한 예측을 하기 위해 사용하는 추가적인 가정

https://re-code-cord.tistory.com/entry/Inductive-Bias%EB%9E%80-%EB%AC%B4%EC%97%87%EC%9D%BC%EA%B9%8C

Inductive Bias



0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114				

- Locality(spatial) Inductive bias

- Translation invariance Inductive bias

Battaglia, Peter W., et al. "Relational inductive biases, deep learning, and graph networks." *arXiv preprint arXiv:1806.01261* (2018).

https://re-code-cord.tistory.com/entry/Inductive-Bias%EB%9E%80-%EB%AC%B4%EC%97%87%EC%9D%BC%EA%B9%8C

Transformer

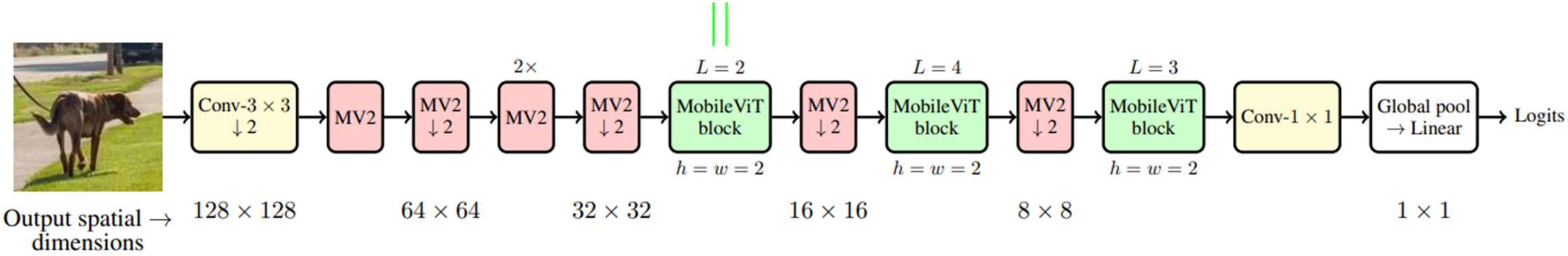
2 _ Method

1. Mobile-ViT Architecture

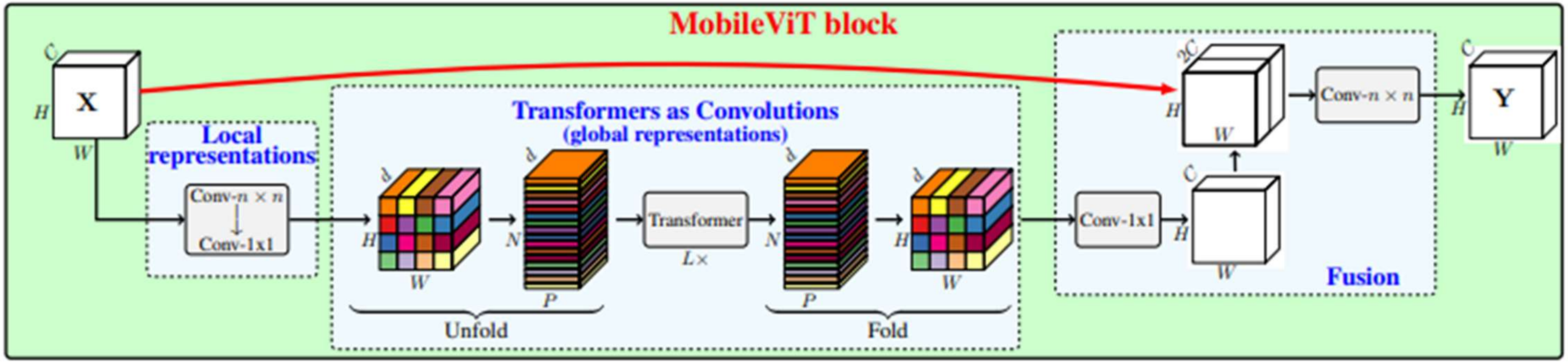
2. Mobile-ViT Block

3. Additional Features

Method
Mobile-ViT Architecture

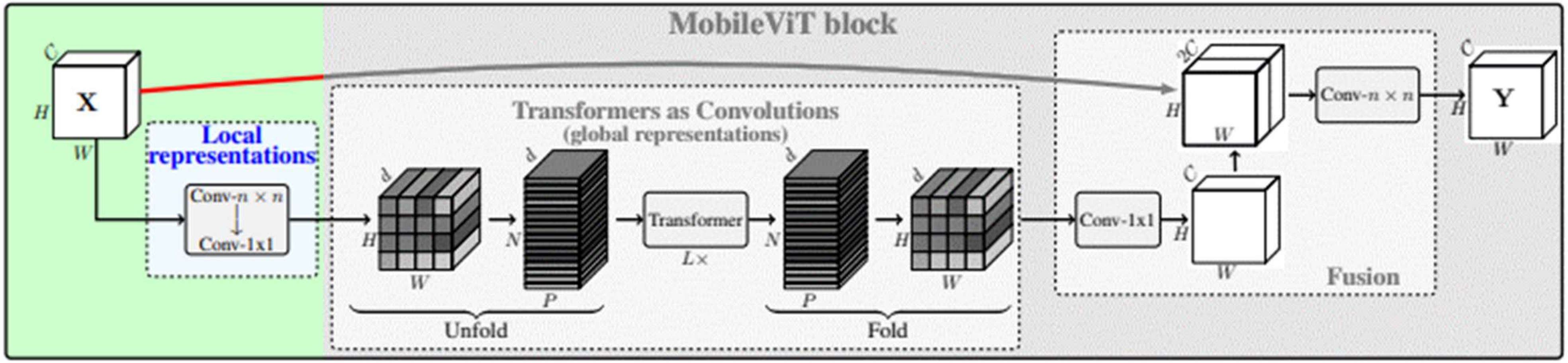


Method
Mobile-ViT Block

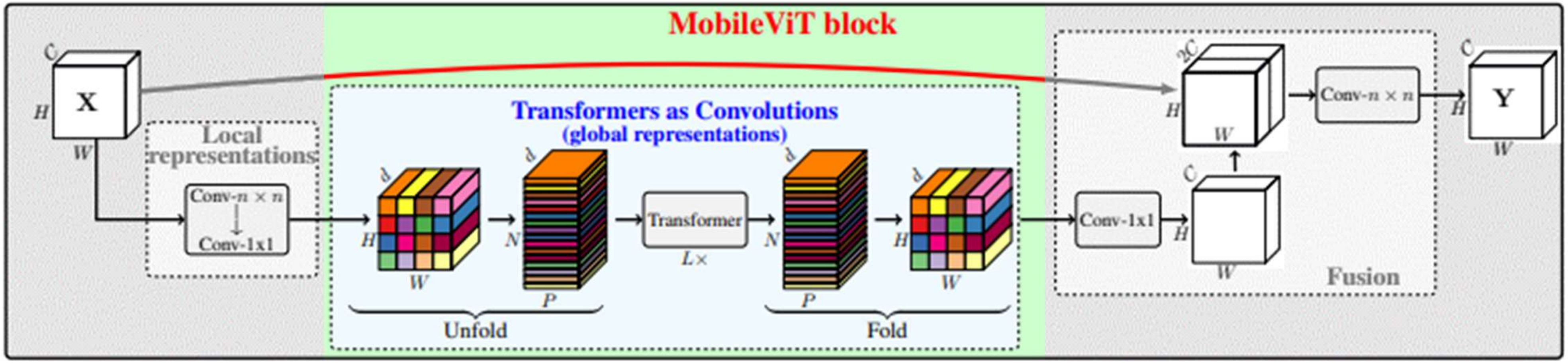


Method

Mobile-ViT Block

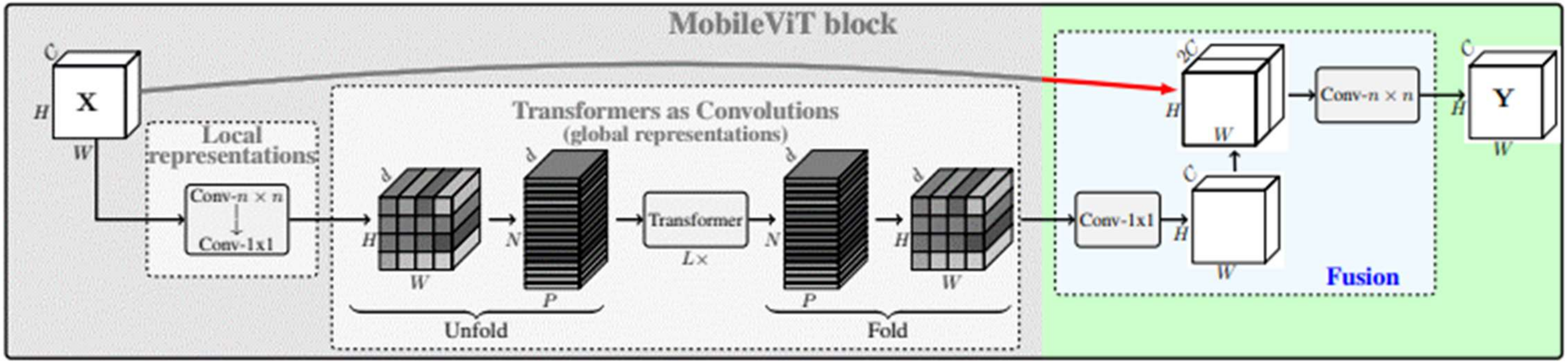


Method
Mobile-ViT Block



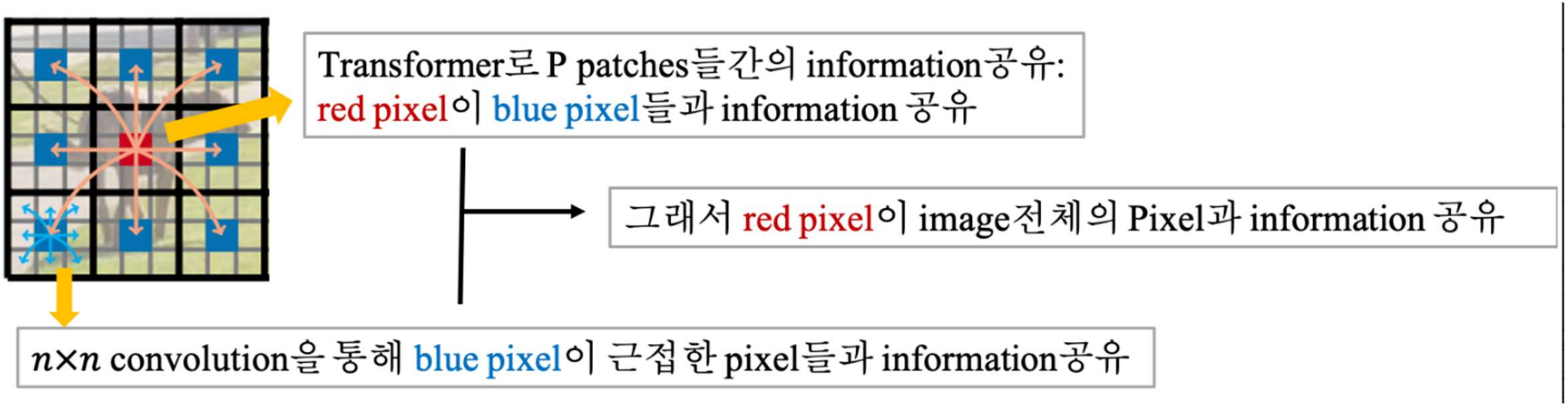
long range non-local dependencies

Method
Mobile-ViT Block



long range non-local dependencies

Method
Mobile-ViT Block



Transformer로 P patches들간의 information공유:
red pixel이 blue pixel들과 information 공유

그래서 red pixel이 image전체의 Pixel과 information 공유

$n \times n$ convolution을 통해 blue pixel이 근접한 pixel들과 information공유

Method
_ Additional Features

- Relationship to convolutions (unfold → matrix multiplication → fold)

- Light weight

- Multi-scale Sampler for Training Efficiency

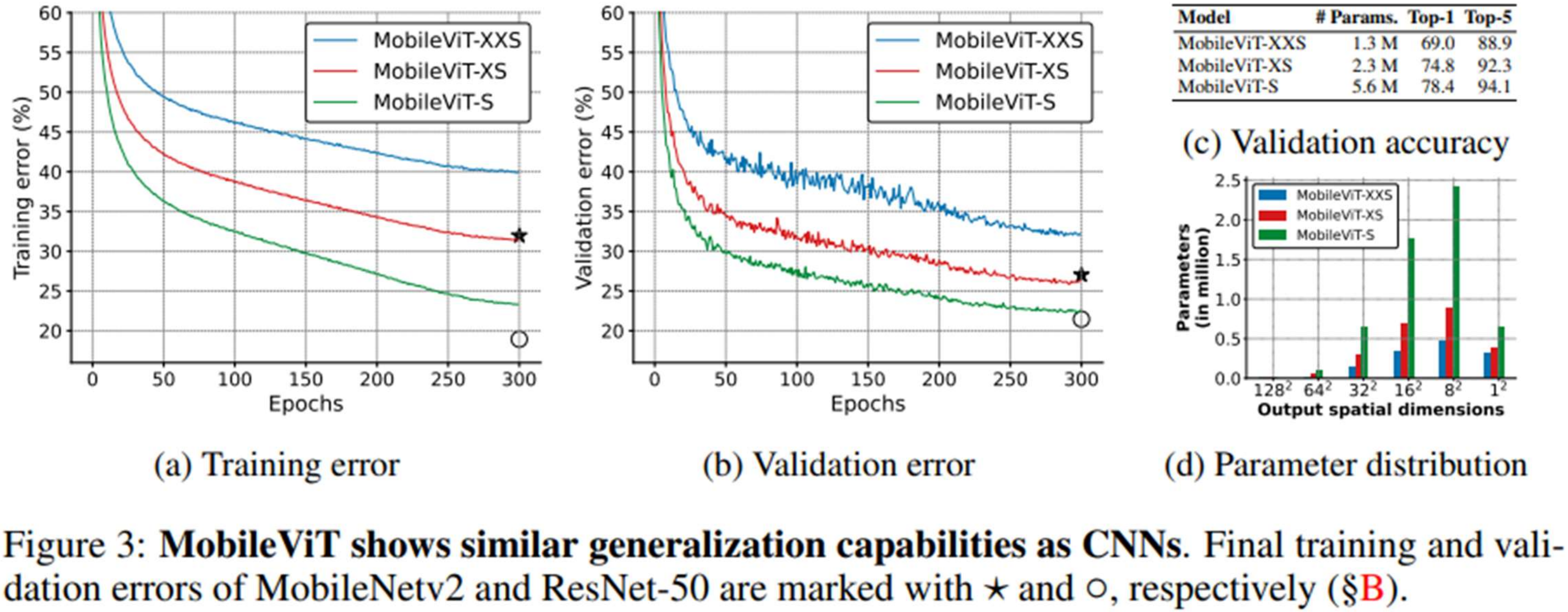
Transformer

3 _ Experiments

1. Model variants
2. Experiments results

Experiments

Model variants



(a) Training error

(b) Validation error

Model	# Params.	Top-1	Top-5
MobileViT-XXS	1.3 M	69.0	88.9
MobileViT-XS	2.3 M	74.8	92.3
MobileViT-S	5.6 M	78.4	94.1

(c) Validation accuracy

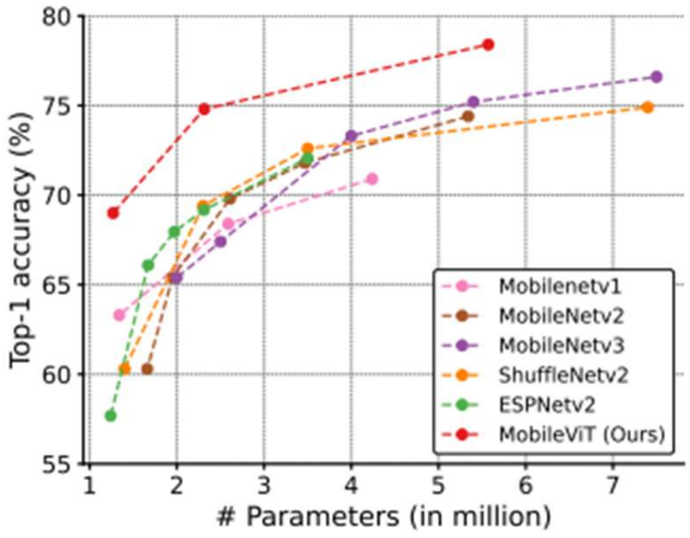
(d) Parameter distribution

Figure 3: **MobileViT shows similar generalization capabilities as CNNs.** Final training and validation errors of MobileNetv2 and ResNet-50 are marked with ★ and ○, respectively (§B).

CHAPTER I _

Experiments

Experiments results



(a) Comparison with light-weight CNNs

Model	# Params. ⇓	Top-1 ⇑
MobileNetv1	2.6 M	68.4
MobileNetv2	2.6 M	69.8
MobileNetv3	2.5 M	67.4
ShuffleNetv2	2.3 M	69.4
ESPNetv2	2.3 M	69.2
MobileViT-XS (Ours)	2.3 M	**74.8**

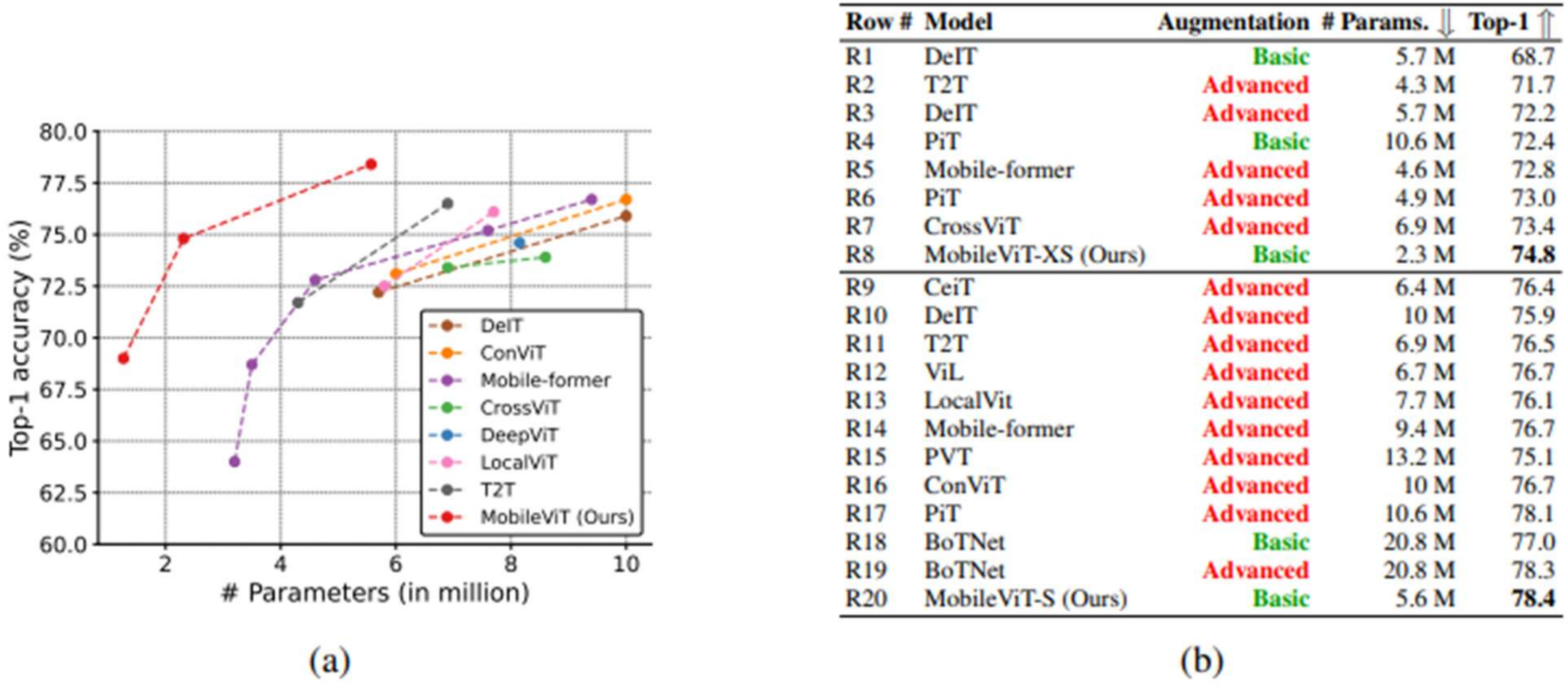
(b) Comparison with light-weight CNNs (similar parameters)

Model	# Params. ⇓	Top-1 ⇑
DenseNet-169	14 M	76.2
EfficientNet-B0	5.3 M	76.3
ResNet-101	44.5 M	77.4
ResNet-101-SE	49.3 M	77.6
MobileViT-S (Ours)	5.6 M	**78.4**

(c) Comparison with heavy-weight CNNs

Figure 6: **MobileViT vs. CNNs** on ImageNet-1k validation set. All models use basic augmentation.

Experiments
_ Experiments results



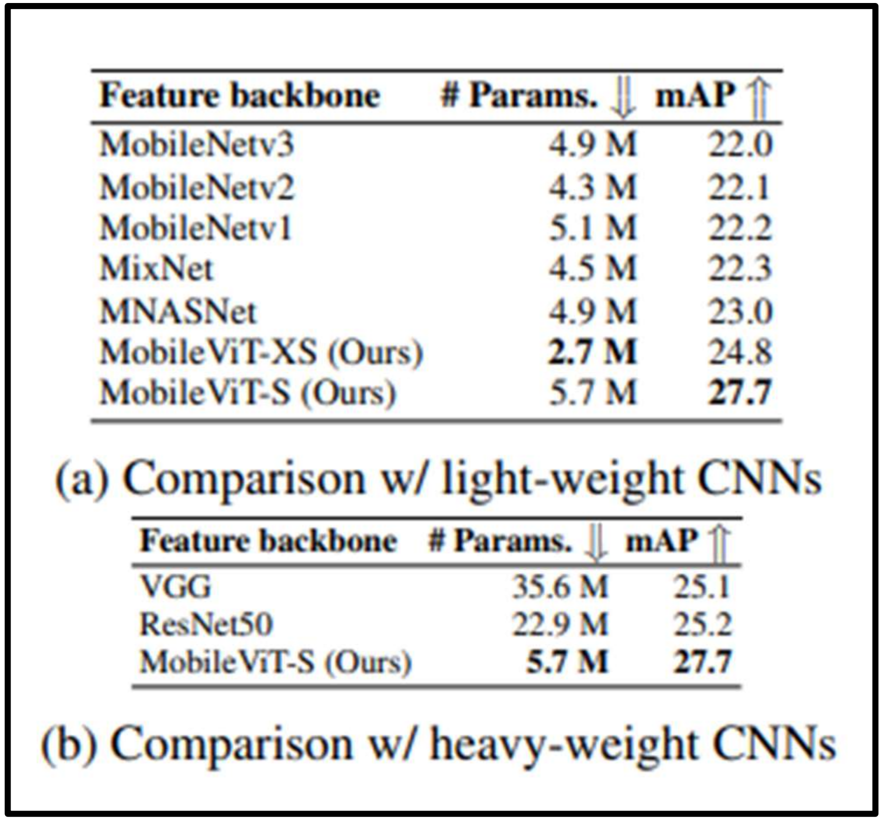
(a)

Row #	Model	Augmentation	# Params. ⇓	Top-1 ⇑
R1	DeIT	Basic	5.7 M	68.7
R2	T2T	Advanced	4.3 M	71.7
R3	DeIT	Advanced	5.7 M	72.2
R4	PiT	Basic	10.6 M	72.4
R5	Mobile-former	Advanced	4.6 M	72.8
R6	PiT	Advanced	4.9 M	73.0
R7	CrossViT	Advanced	6.9 M	73.4
R8	MobileViT-XS (Ours)	Basic	2.3 M	**74.8**
R9	CeiT	Advanced	6.4 M	76.4
R10	DeIT	Advanced	10 M	75.9
R11	T2T	Advanced	6.9 M	76.5
R12	ViL	Advanced	6.7 M	76.7
R13	LocalVit	Advanced	7.7 M	76.1
R14	Mobile-former	Advanced	9.4 M	76.7
R15	PVT	Advanced	13.2 M	75.1
R16	ConViT	Advanced	10 M	76.7
R17	PiT	Advanced	10.6 M	78.1
R18	BoTNet	Basic	20.8 M	77.0
R19	BoTNet	Advanced	20.8 M	78.3
R20	MobileViT-S (Ours)	Basic	5.6 M	**78.4**

(b)

Figure 7: **MobileViT vs. ViTs** on ImageNet-1k validation set. Here, basic means ResNet-style augmentation while **advanced** means a combination of augmentation methods with basic (e.g., MixUp (Zhang et al., 2018), RandAugmentation (Cubuk et al., 2019), and CutMix (Zhong et al., 2020)).

Experiments results



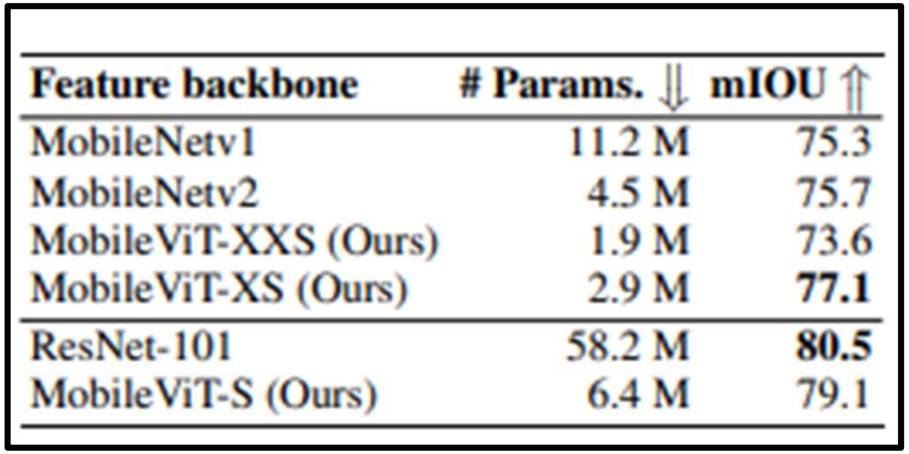
Feature backbone	# Params. ⇓	mAP ⇑
MobileNetv3	4.9 M	22.0
MobileNetv2	4.3 M	22.1
MobileNetv1	5.1 M	22.2
MixNet	4.5 M	22.3
MNASNet	4.9 M	23.0
MobileViT-XS (Ours)	**2.7 M**	24.8
MobileViT-S (Ours)	5.7 M	**27.7**

(a) Comparison w/ light-weight CNNs

Feature backbone	# Params. ⇓	mAP ⇑
VGG	35.6 M	25.1
ResNet50	22.9 M	25.2
MobileViT-S (Ours)	**5.7 M**	**27.7**

(b) Comparison w/ heavy-weight CNNs

Detection



Feature backbone	# Params. ⇓	mIOU ⇑
MobileNetv1	11.2 M	75.3
MobileNetv2	4.5 M	75.7
MobileViT-XXS (Ours)	1.9 M	73.6
MobileViT-XS (Ours)	2.9 M	**77.1**
ResNet-101	58.2 M	**80.5**
MobileViT-S (Ours)	6.4 M	79.1

Segmentation

감사합니다

Thank you

2022. 10 24