Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

School of Industrial and Management Engineering, Korea University

Lee Kyung Yoo

KOREA UNIVERSITY

DMQA

Contents

Introduction

❖ Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet (arXiv, 2021)

- National University of Singapore과 YITU Technology에서 연구
- 2021년 11월 25일 기준으로 193회 인용

Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

Li Yuan[1]*, Yunpeng Chen[2], Tao Wang[1]*, Weihao Yu[1], Yujun Shi[1],
Zihang Jiang[1], Francis E.H. Tay[1], Jiashi Feng[1], Shuicheng Yan[2]

[1]National University of Singapore [2] YITU Technology
yuanli@u.nus.edu, yunpeng.chen@yitu-inc.com, shuicheng.yan@gmail.com

Abstract

Transformers, which are popular for language modeling, have been explored for solving vision tasks recently, e.g., the Vision Transformer (ViT) for image classification. The ViT model splits each image into a sequence of tokens with fixed length and then applies multiple Transformer layers to model their global relation for classification. However, ViT achieves inferior performance to CNNs when trained from scratch on a midsize dataset like ImageNet. We find it is because: 1) the simple tokenization of input images fails to model the important local structure such as edges and lines among neighboring pixels, leading to low training sample efficiency; 2) the redundant attention backbone design of ViT leads to limited feature richness for fixed computation budgets and limited training samples. To overcome such limitations, we propose a new Tokens-To-Token Vision Transformer (T2T-ViT), which incorporates 1) a layer-wise Tokens-to-Token (T2T) transformation to progressively structurize the image to tokens by recursively aggregating neighboring Tokens into one Token (Tokens-to-Token), such that local structure represented by surrounding tokens can be modeled and tokens length can be reduced; 2) an efficient backbone with a deep-narrow structure for vision transformer motivated by CNN architecture design after empirical study. Notably, T2T-ViT reduces the parameter count and MACs of vanilla ViT by half, while achieving more than 3.0% improvement when trained from scratch on ImageNet. It also outperforms ResNets and achieves comparable performance with MobileNets by directly training on ImageNet. For example, T2T-ViT with comparable size to ResNet50 (21.5M parameters) can achieve 83.3% top1 accuracy in image resolution 384×384 on ImageNet. [1]

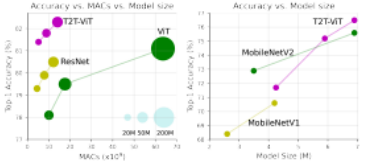


Figure 1. Comparison between T2T-ViT with ViT, ResNets and MobileNets when trained from scratch on ImageNet. Left: performance curve of MACs vs. top-1 accuracy. Right: performance curve of model size vs. top-1 accuracy.

tion [4, 62] and image processing like denoising, super-resolution and deraining [5]. Among them, the Vision Transformer (ViT) [14] is the first full-transformer model that can be directly applied for image classification. In particular, ViT splits each image into 14×14 or 16×16 patches (*a.k.a.*, tokens) with fixed length; then following practice of the transformer for language modeling, ViT applies transformer layers to model the global relation among these tokens for classification.

Though ViT proves the full-transformer architecture is promising for vision tasks, its performance is still inferior to that of similar-sized CNN counterparts (*e.g.* ResNets) when trained from scratch on a midsize dataset (*e.g.*, ImageNet). We hypothesize that such performance gap roots in two main limitations of ViT: 1) the straightforward tokenization of input images by hard split makes ViT unable to model the image local structure like edges and lines, and thus it requires significantly more training samples (like JFT-300M for pretraining) than CNNs for achieving similar

DMQA

Introduction

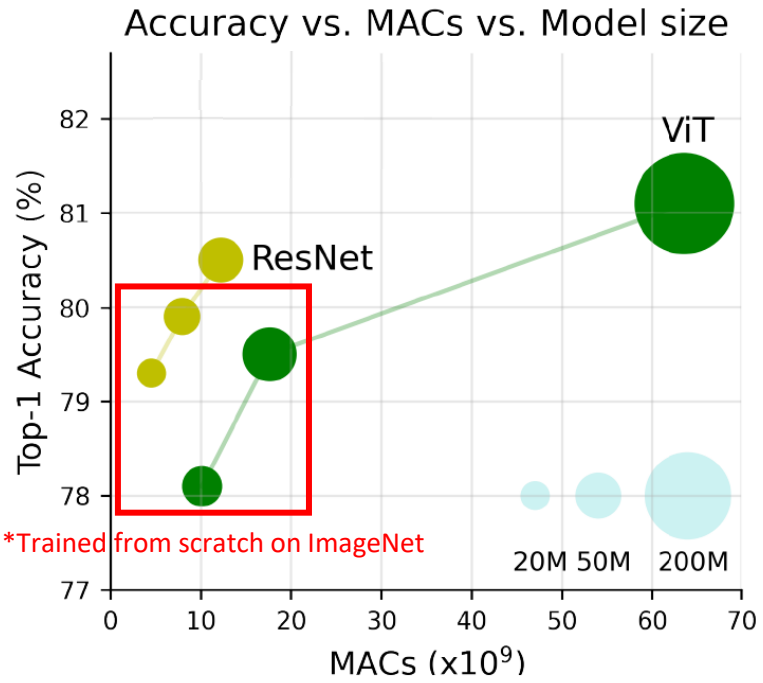
❖ Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet (arXiv, 2021)

- Image classification 분야에서 midsize dataset으로 pre-training 없이 학습하는 경우, 비슷한 크기의 CNN보다 우세한 성능을 보이는 ViT architecture를 새롭게 제안
- 기존 ViT 구조와 비교하여 크게 두 가지 구조적 차이 존재
  - ➢ 이미지 구조 정보 파악에 보다 적합한 **Tokens-to-Token (T2T) module**
  - ➢ Feature richness 향상을 이루는 CNN 기반의 **Deep-narrow backbone (T2T-ViT) architecture**
- 기존 ViT와 비교 시 학습 파라미터와 연산량을 절반으로 감소시켰으며, 동시에 ImageNet으로 pre-training 없이 학습한 경우에도 3% 향상된 성능을 보임
- CNN 기반의 모델과 비교 시 ResNets보다 향상된 성능, MobileNets에 준하는 성능을 보임

Research Purpose

❖ Midsize dataset으로 학습 시 CNN보다 낮은 성능의 ViT

- ImageNet과 같은 midsize dataset을 통해 random initialized된 가중치로부터 모델을 학습할 시, 비슷한 크기의 CNN(e.g. ResNets)에 비해 ViT의 성능이 여전히 떨어짐



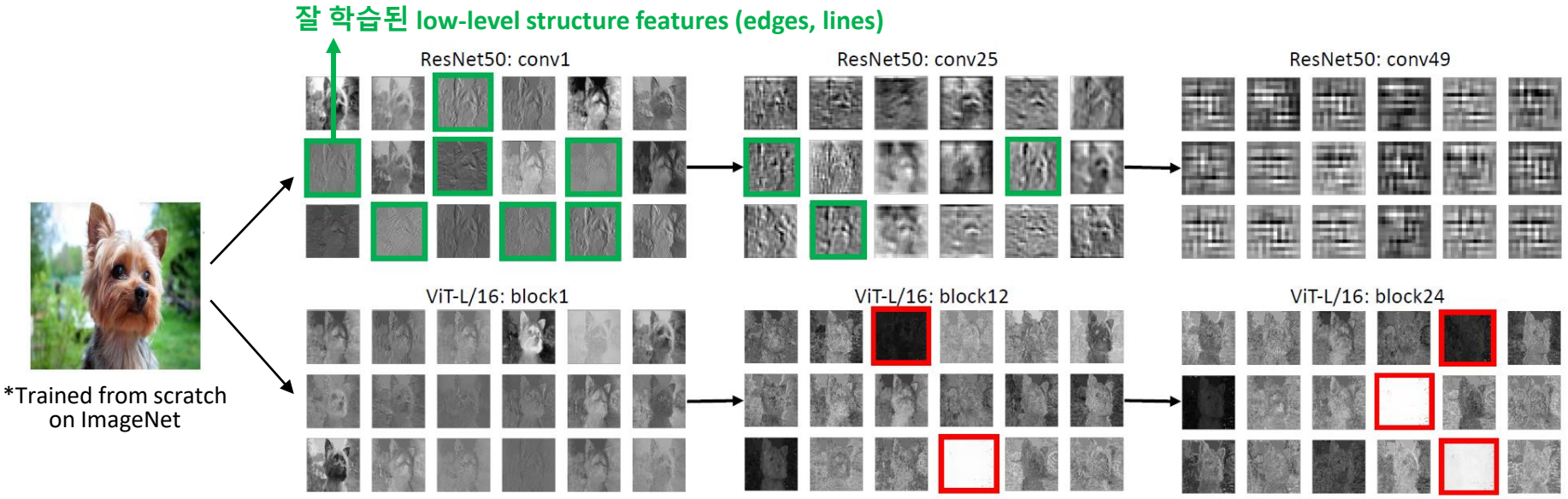
DMQA

Research Purpose

❖ 이러한 성능 저하의 원인으로 가정한 ViT의 한계점

1) 입력 이미지의 단순 토큰화

- 인접 픽셀 간의 edges, lines 등 중요한 local structure를 모델링하지 못하는데 영향을 미침
- ResNet50과 ViT-L/16로부터 학습된 features를 시각화하여 비교한 결과, ViT-L/16 의 경우 global structure는 잘 학습되는 반면 local structure는 잘 학습되지 못함



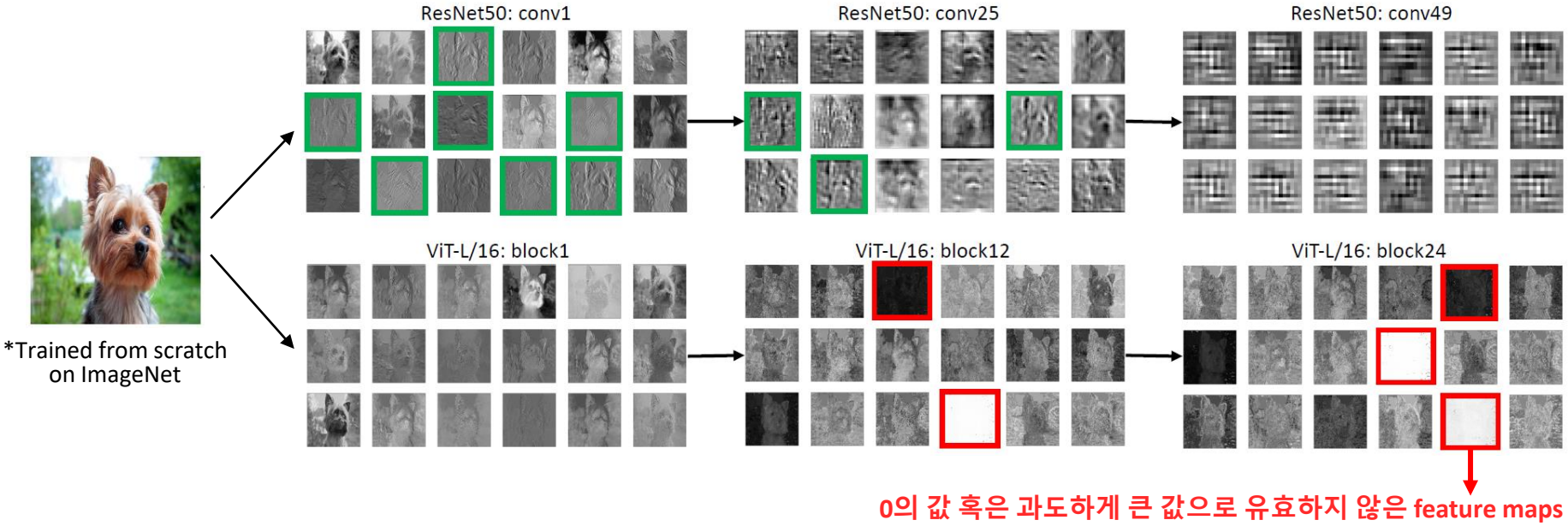
DMQA

Research Purpose

❖ 이러한 성능 저하의 원인으로 가정한 ViT의 한계점

2) ViT backbone의 자체적 결함

- 반복적인 attention backbone이 feature richness를 충분히 향상시키지 못하는데 영향을 미침
- ResNet50과 ViT-L/16로부터 학습된 features를 시각화하여 비교한 결과, ViT-L/16 의 경우 유효한 feature maps를 형성하지 못하는 경우 존재함



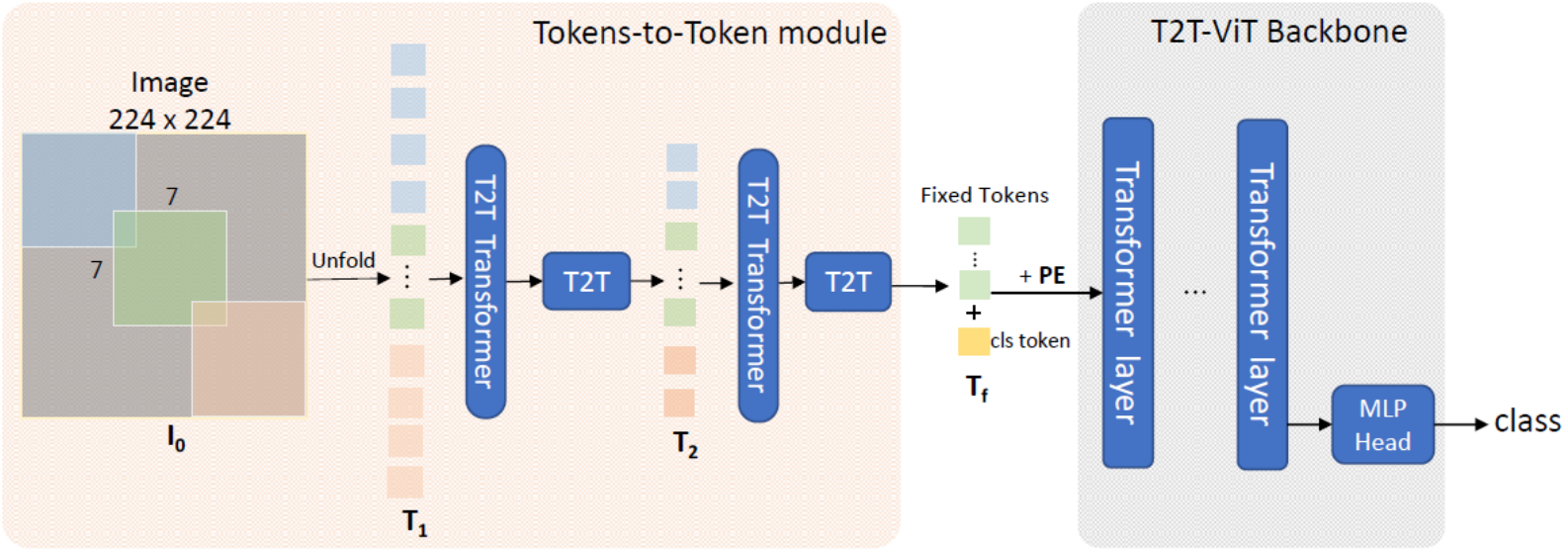
DMQA

Tokens-to-Token ViT

- Overview of Tokens-to-Token ViT

❖ Architecture

- 한계점 1) 입력 이미지의 단순 토큰화 → 이미지의 구조적 형태를 보존 가능하도록 토큰화
  - ➢ **Tokens-to-Token (T2T) module**
- 한계점 2) ViT backbone의 자체적 결함 → CNN 기반의 효율적인 backbone으로 재설계
  - ➢ **Deep-narrow backbone (T2T-ViT) architecture**

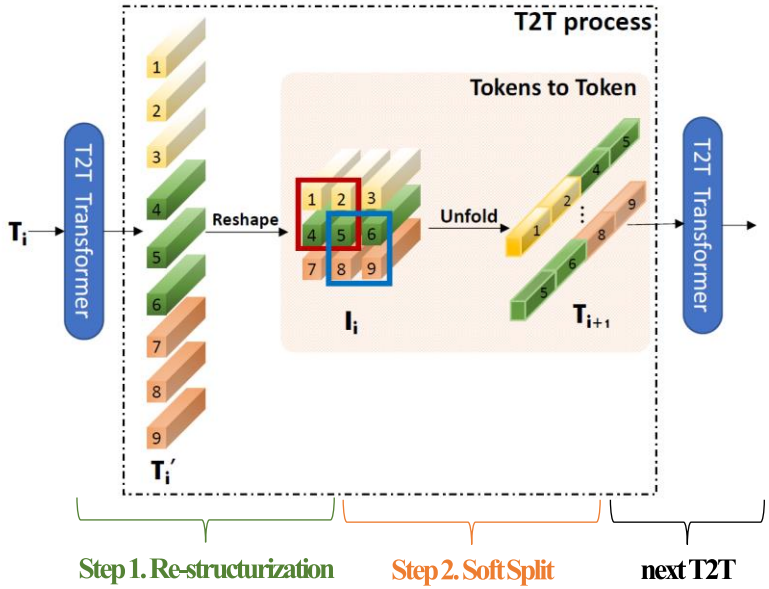


Tokens-to-Token ViT

- Tokens-to-Token (T2T) module

❖ Architecture

- 각 T2T module은 두 단계로 구성
  - ➢ **Step 1. Re-structurization**: Spatial한 이미지 형상처럼 **토큰을 reshape**하는 과정
  - ➢ **Step 2. Soft Split**: Reshape한 이미지를 overlapping하여 **각 패치들로 split**하는 과정
- 각 T2T module을 통해 전반적으로 이미지의 local structure information을 잘 학습할 수 있으며, 토큰의 length(개수)를 반복적으로 줄일 수 있음



$$T_i' = MLP\big(MSA(T_i)\big)$$

$$I_i = Reshape(T_i')$$

$$T_{i+1} = SS(I_i)$$

$$T_1 = SS(I_0)$$

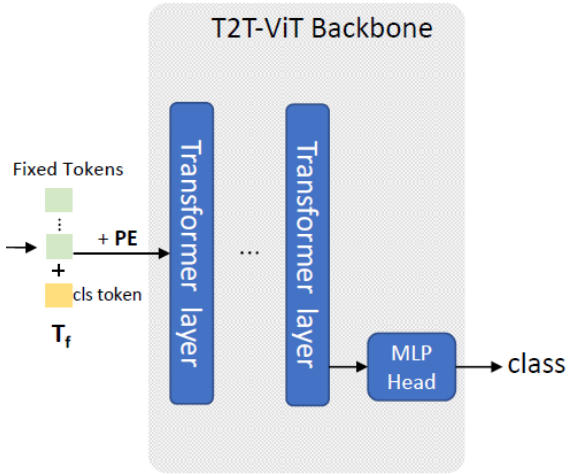
$$(i = 1, \dots, n-1)$$

Tokens-to-Token ViT

- Deep-narrow backbone (T2T-ViT) architecture

❖ Architecture

- Vision transformer의 효율적인 backbone을 찾기 위해 5가지 종류의 CNN architecture를 바탕으로 transformer layer를 어떠한 구조로 구성할지 비교 실험 진행
- **ResNet 기반 deep-narrow architecture(ViT-DN)**를 적용했을 시 가장 좋은 성능을 보임
- ViT-DN은 384 hidden dimensions로 구성된 16개의 transformer layer로 이루어짐



적용 모델

1) **DenseNet** 기반 dense connection
2) **ResNet** 기반 deep-narrow architecture
   Wide-ResNet 기반 shallow-wide architecture
3) **SENet** 기반 channel attention
4) **ResNeXt** 기반 more split heads in multi-head attention layer
5) **GhostNet** 기반 ghost operation

Tokens-to-Token ViT

- Deep-narrow backbone (T2T-ViT) architecture

❖ Architecture

- 결과적으로 구성한 T2T-ViT backbone은 ViT와 동일하게 T2T module로부터의 고정 길이의 토큰을 입력으로 사용함
- 그러나 **ViT보다 더 깊고 좁은 구조를 적용**
  - ➢ T2T-ViT-14는 384 hidden dimensions로 구성된 14개의 transformer layer로 이루어짐
  - ➢ ViT-B/16는 768 hidden dimensions로 구성된 12개의 transformer layer로 이루어짐
  - ➢ 파라미터수와 연산량을 3배 가량 감소시켜 효율적인 구조를 이룸
- Channel dimensions 감소 → Channel 간 redundancy 감소 (= 중복된 feature 수 감소) → Feature richness 향상으로 이어져 성능 향상까지 이룸

Experiments

❖ Architecture for model comparisons

- 모델 사이즈 비교를 위한 ViT / T2T-ViT structure details

Models	Tokens-to-Token module				T2T-ViT backbone			Model size	
	T2T transformer	Depth	Hidden dim	MLP size	Depth	Hidden dim	MLP size	Params (M)	MACs (G)
ViT-S/16 [14]	-	-	-	-	8	786	2358	48.6	10.1
ViT-B/16 [14]	-	-	-	-	12	786	3072	86.8	17.6
ViT-L/16 [14]	-	-	-	-	24	1024	4096	304.3	63.6
T2T-ViT-14	Performer	2	64	64	14	384	1152	21.5	5.2
T2T-ViT-19	Performer	2	64	64	19	448	1344	39.2	8.9
T2T-ViT-24	Performer	2	64	64	24	512	1536	64.1	14.1
T2T-ViT$_t$-14	Transformer	2	64	64	14	384	1152	21.5	6.1
T2T-ViT-7	Performer	2	64	64	8	256	512	4.2	1.2
T2T-ViT-12	Performer	2	64	64	12	256	512	6.8	2.2

Experiments

❖ Default hyperparameter settings

- 사용 데이터셋
  - ➢ ImageNet dataset (1.3 million images in training set / 50k images in validation set)

Models	T2T-ViT-7/12	T2T-ViT-14	T2T-ViT-19/24
Epochs	310	310	310
Warmup Epochs	5	5	5
Batch size	1024	512	512
Learning rate	1e-3	5e-4	5e-4
Weight decay	3e-2	5e-2	6.5e-2
Label smoothing	0.1	0.1	0.1
Dropout	0	0	0
Stoch.Depth	0.1	0.1	0.1
Mixup prob.	0.8	0.8	0.8
Cutmix prob.	1.0	1.0	1.0
Erasing prob.	0.25	0.25	0.25

Experiments

❖ Comparisons with ViT

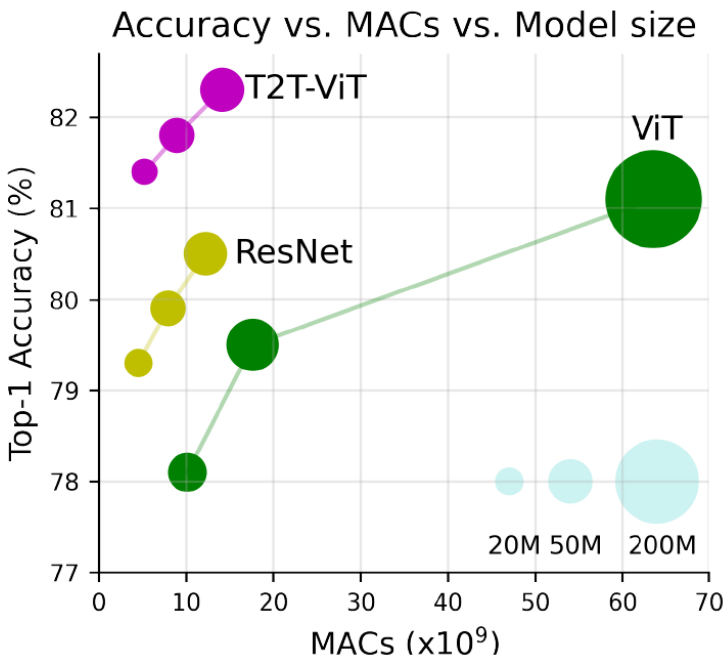
- T2T-ViT는 파라미터수와 연산량에서 ViT보다 훨씬 작지만 더 높은 성능을 보임

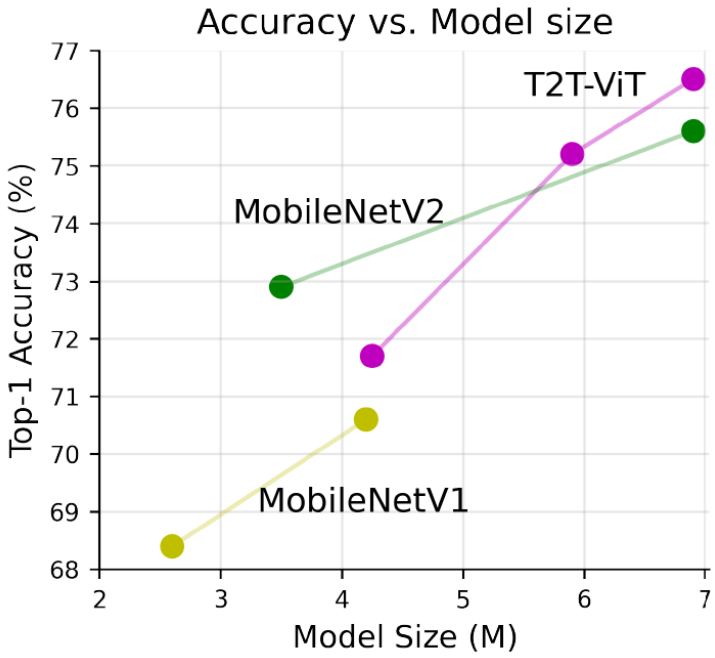
Models	Top1-Acc (%)	Params (M)	MACs (G)
ViT-S/16 [14]	78.1	48.6	10.1
DeiT-small [38]	79.9	22.1	4.6
DeiT-small-Distilled [38]	81.2	22.1	4.7
T2T-ViT-14	**81.5**	21.5	5.2
T2T-ViT-14↑384	**83.3**	21.5	17.1
ViT-B/16 [14]	79.8	86.4	17.6
ViT-L/16 [14]	81.1	304.3	63.6
T2T-ViT-24	**82.3**	**64.1**	**14.1**

Experiments

❖ Comparisons with ResNets / MobileNets

- T2T-ViT는 전반적으로 ResNet과 같이 중간 크기의 모델일 때 우수한 성능을 얻을 수 있고, MobileNet과 같이 작은 크기의 모델일 때 기존 결과에 견줄 만한 합리적인 성능 도출 가능





DMQA

Experiments

❖ From CNN to ViT

- SE(ViT-SE)와 Deep-Narrow (ViT-DN) 구조 모두 ViT에 이점이 있음을 확인 가능
- CNN으로부터 ViT에 적용될 만한 가장 효과적인 구조는 모델 크기와 MAC을 약 2배 줄이고, 0.9%의 성능 개선을 이룬 deep-narrow 구조

Model Type	Models	Top1-Acc (%)	Params (M)	MACs (G)	Depth	Hidden_dim
CNN to ViT	ViT-S/16 (Baseline)	78.1	48.6	10.1	8	768
	ViT-DN	79.0 (+0.9)	24.5	5.5	16	384
	ViT-SW	69.9 (-8.2)	47.9	9.9	4	1024
	ViT-Dense	76.8 (-1.3)	46.7	9.7	19	128-736
	ViT-SE	**78.4** (+0.3)	49.2	10.2	8	768
	ViT-ResNeXt	78.0 (-0.1)	48.6	10.1	8	768
	ViT-Ghost	73.7 (-4.4)	32.1	6.9	8	768
CNN to T2T-ViT	T2T-ViT-14 (Baseline)	81.5	21.5	5.2	14	384
	T2T-ViT-Wide	77.9 (-3.4)	25.1	5.4	14	768
	T2T-ViT-Dense	80.6 (-1.1)	23.7	5.9	19	128-584
	T2T-ViT-SE	**81.6** (+0.1)	21.9	5.2	14	384
	T2T-ViT-ResNeXt	81.5 (+0.0)	21.5	5.2	14	384
	T2T-ViT-Ghost	79.5 (-2.0)	16.3	3.7	14	384

Conclusion

❖ Conclusion

- Transformer architecture를 재설계함으로써 ViT가 JFT-300M에 대한 pre-training 없이 다양한 조건 하에서 CNN보다 좋은 성능을 이루어 낼 수 있음을 최초로 입증
- 본 논문에서 제안한 T2T-ViT는 ViT의 단순한 토큰화 방식보다 더 효과적인 토큰화 방법인 T2T module을 통해 각 토큰에 대한 중요한 local structure를 인코딩 가능
- 이와 함께 다양한 비교 실험을 통해 efficient backbone을 구성함으로써 중복성 감소 및 feature richness 향상, 결과적으로 deep-narrow architecture가 ViT에 가장 적합함을 발견

References

- Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
- Bhojanapalli, Srinadh, et al. "Leveraging redundancy in attention with Reuse Transformers." arXiv preprint arXiv: 2110.06821 (2021).
- https://junha1125.github.io/blog/artificial-intelligence/2021-03-25-T2T/
- https://www.youtube.com/watch?v=eaZt9asVYH0


Thank You