PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

School of Industrial and Management Engineering, Korea University

Sae Rin Lim

KOREA UNIVERSITY

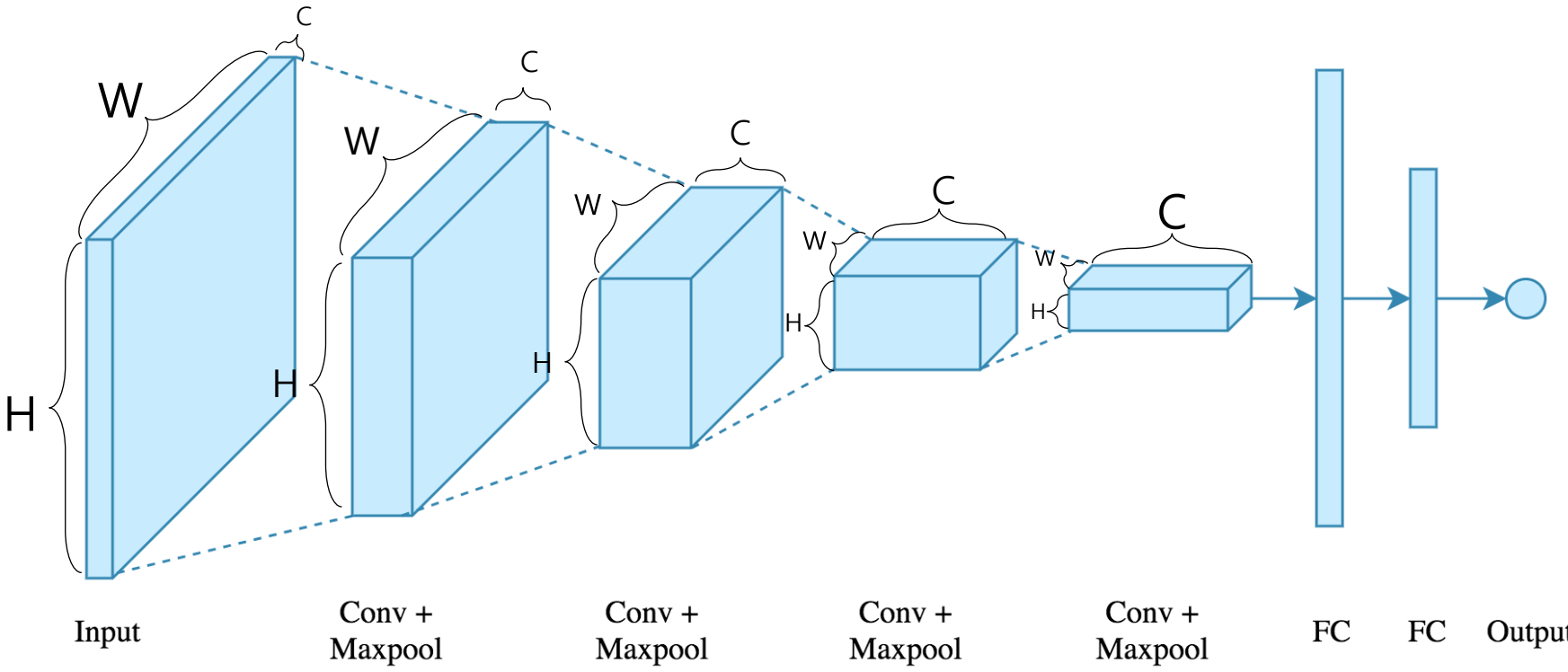
DMQA

• Contents

• Research Purpose

❖ **Two levels of Redundancies in the ViT, (1) Fixed number of tokens**

- CNN에서 층을 깊게 쌓을수록 High-level 특징을 캐치하는 경향이 있음
- CNN은 더 의미 있는 high-level 특징을 추출하기 위해 층이 깊어질수록 해상도(W, H)를 줄이는 동시에 채널(C)을 늘리도록 디자인 되어 있음

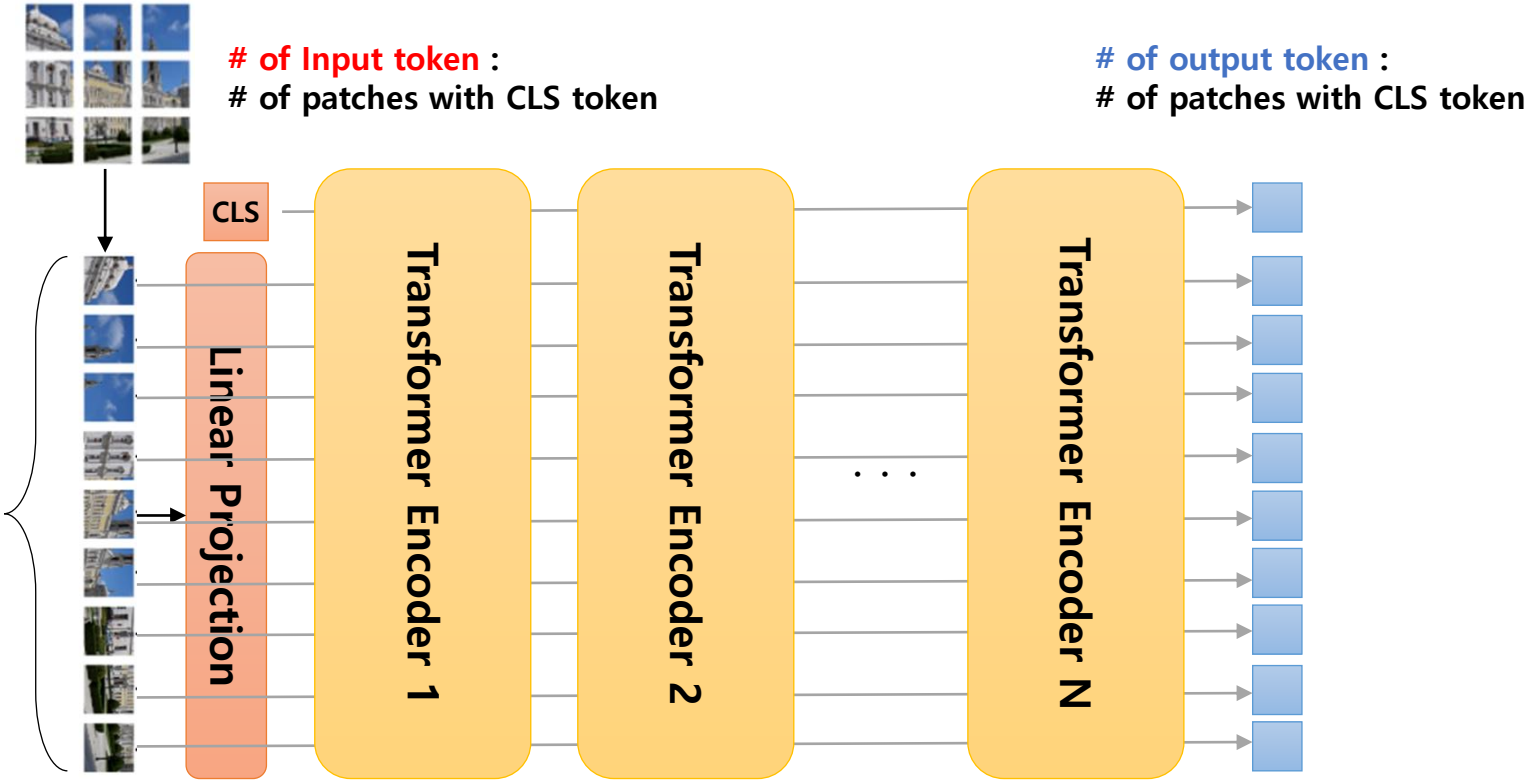


https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

• Research Purpose

❖ **Two levels of Redundancies in the ViT, (1) Fixed number of tokens**

- Vision Transformer는 이러한 디자인이 반영되어 있지 않고 고정된 토큰의 수를 사용
- 이 고정된 토큰 수는 low-level 특징을 캐치하는 것에 충분하지 못할 가능성이 있으며, **중복되는 high-level 특징을 추출하는 것을 발견**

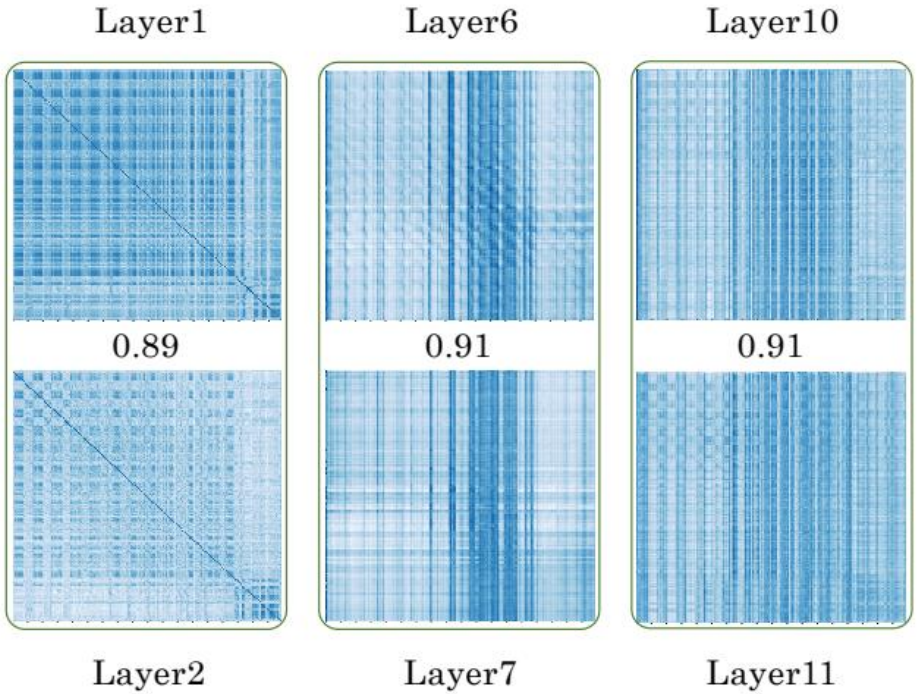


DMQA

• Research Purpose

❖ **Two levels of Redundancies in the ViT, (2) Attention maps between adjacent encoders**

- 인접한 Transformer encoder 사이에서 생성된 **attention map이 유사하다는 것을 발견**



• Introduction to PSViT

❖ PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- 2021년 8월 7일 arXiv에 공개
- 2021년 09월 10일 기준 0회 인용
- 앞선 두 문제를 해결하기 위해 **Token Pooling과 Attention Sharing를 활용**하여 성능을 높이고 계산 효율성을 증가

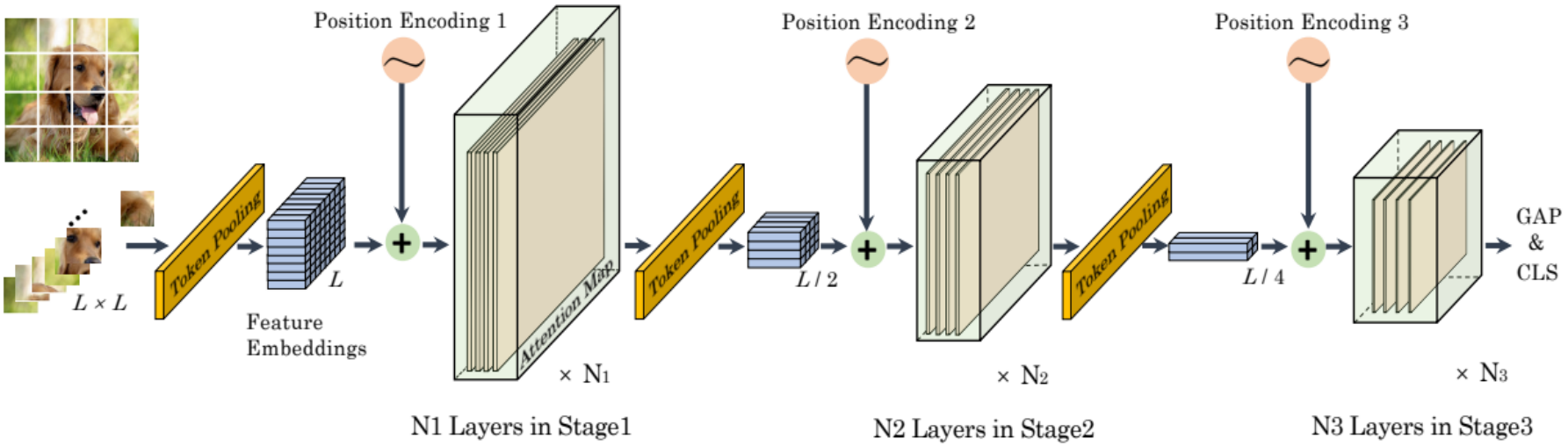
PSViT: Better Vision Transformer via Token Pooling and Attention Sharing

Boyu Chen[1]*, Peixia Li[1]*, Baopu Li[2]*, Chuming Li[3], Lei Bai[1], Chen Lin[4],
Ming Sun[3], Junjie Yan[3], Wanli Ouyang[1]
[1] The University of Sydney, [2] BAIDU USA LLC,
[3] SenseTime Group Limited, [4] University of Oxford

• Introduction to PSViT

❖ PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- 전체적인 Framework는 총 3단계로 나뉘어져 있음(단계 별로 Attention size를 맞추기 위해)
- Stage가 지남에 따라 토큰의 수는 L*L개에서 L*L/16개로 줄어들며 토큰의 차원(feature embeddings)은 늘어남
- 마지막 층에서 생성된 feature에 Global Average Pooling을 적용하여 최종적인 feature vector 생성



DMQA

• Token pooling

❖ PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

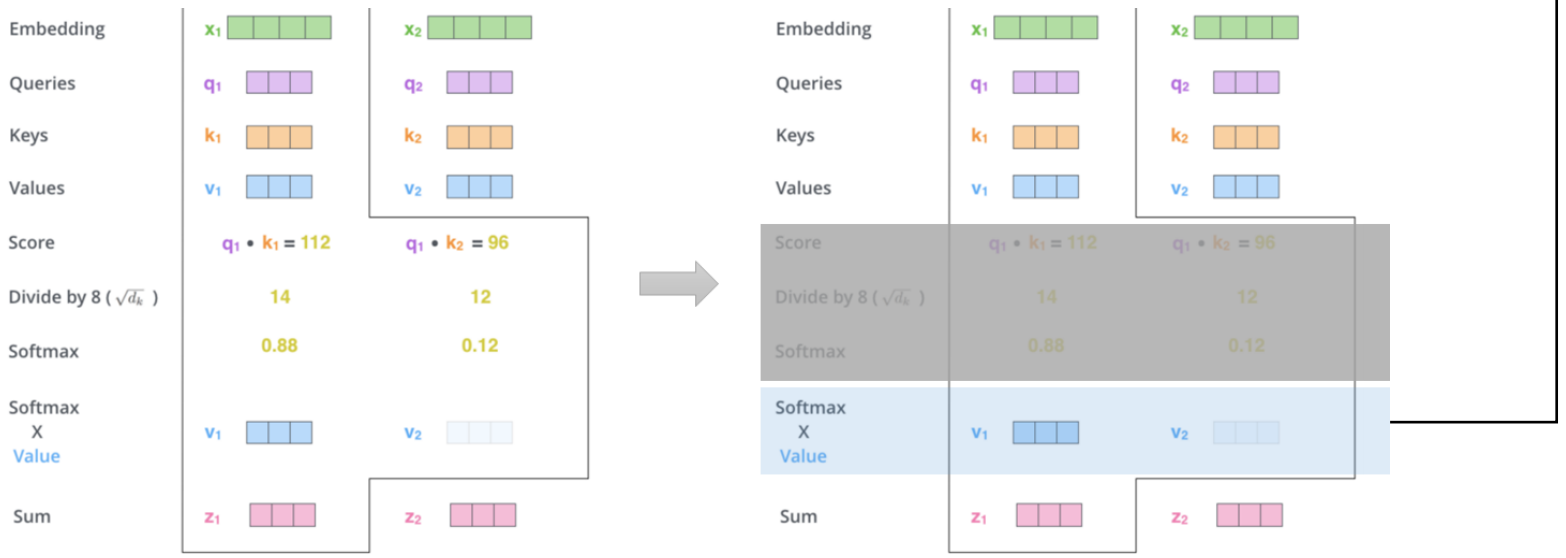
- 토큰의 개수를 줄이기 위한 방법으로 pooling을 사용
- 같은 연산량(FLOPS 기준)을 유지하면서 토큰의 수를 줄이면서 단계별 인코더를 늘리는 방법(Token Dimension1)과 토큰의 차원을 늘리는 방법(Token Dimension1)을 비교 실험 후 후자 선택
- 이러한 pooling층은 CNN처럼 layer마다 진행하는 것이 아닌 각 단계 별로 진행
- 토큰의 형태가 **1D array인 PSViT-1D**는 작은 Kernel size를 가진 conv 1D로 구성하여 다운샘플링
- 토큰의 형태가 **2D array인 PSViT-2D**는 conv 2D에 stride를 2로 설정하여 다운샘플링

Model	DeiT-Tiny	Token Dimension1	Token Dimension2
w/o Pooling	no Pooling	2 Poolings	2 Poolings
Token Dimensions	[192, 192, 192]	[192, 192, 192]	[192, 256, 384]
Token Numbers	[197, 197, 197]	[197, 99, 50]	[197, 99, 50]
Layer Numbers	[4, 4, 4]	[4, 8, 20]	[4, 4, 4]
FLOPS(G)	1.3	1.3	1.3
Top1 Acc	72.2%	75.0%	76.3 %

• Attention Sharing

❖ PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- 인접한 층에서 입력되는 특징이 residual connection에 의해 스무스하게 변화하기 때문에 attention map 이 유사한 것이라고 해석
- Sharing mechanism을 적용하여 이미 계산한 attention score를 재사용한다면 **Value matrix와의 내적**만으로 계산되어 redundancy를 줄이는 동시에 연산량을 줄일 수 있음
- 반면 다른 특징을 추출하는 head로의 공유는 오히려 악영향을 끼치기 때문에 이를 옵션으로 둠



Original attention 연산 과정

Attention score를 공유 받는 인코더의 연산 과정

DMQA

• Auto ML for hyper-parameters

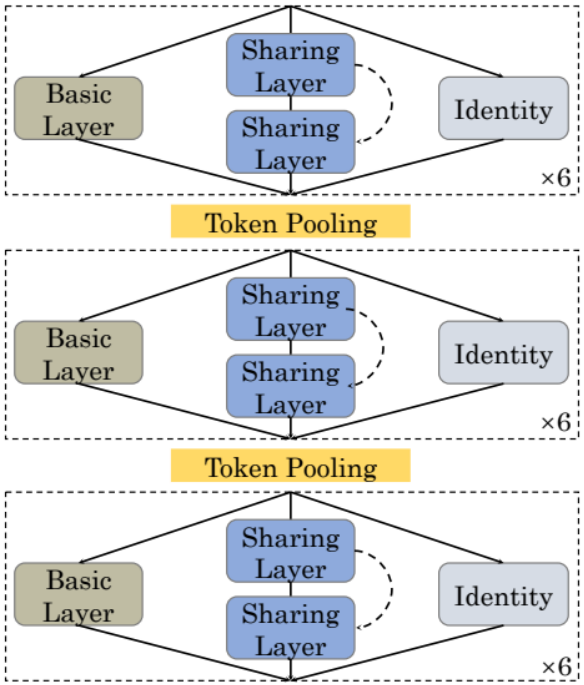
❖ PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- PSViT에서 모델 구조를 결정하는 주요 hyper-parameter는 토큰의 수($number\ of\ token : N_t$), 토큰 하나의 차원($feature\ dimension : N_f$), 단계별 인코더의 수($number\ of\ layers\ for\ each\ stage : N_s$)로 구성
- 최적의 모델구조를 찾기 위해서는 층마다 $(N_t * N_f * N_s)$ 이라는 Search space를 탐색해야 함
- 이 방대한 Search space를 줄이기 위해 아래와 같은 제약을 줌
  - Token pooling에서의 비교 실험을 통해 토큰 차원이 늘어나면 토큰 수를 줄이는 경우의 수만 탐색
  - 같은 단계에서는 토큰의 수를 같게 고정하였으며 3 단계로 고정
  - 단계별 인코더의 수의 최대값에 제한을 둠
  - Attention sharing을 사용할지 선택할 수 있도록 함

• Auto ML for hyper-parameters

❖ PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- 빠른 탐색을 위해 유전적 알고리즘이나 강화학습을 통한 Auto ML이 아닌 weight sharing based method를 사용하였으며 이를 위해 모든 후보 아키텍처를 포함하는 **supernet** 정의
- Supernet은 3단계으로 구성되어 있으며 각 단계 별로 6개의 셀(=최대 인코더 수)가 존재
- 각각의 셀은 일반적인 **Basic layer**, **Attention sharing mechanism**이 적용된 **sharing layer**, 그리고 **이전 값을 그대로 사용하는 Identity(encoder가 필요없다는 선택지)**, 총 세 가지 중 하나를 선택



Supernet of PSViT

• Experiments

PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- ImageNet 데이터셋을 통한 Classification 성능 비교
- *는 같은 세팅(batch size, optimizer 등)에서의 실험 결과를 의미

Model	FLOPS(G)	Acc(%)
ResNet18	1.8	69.8
ResNet18*	1.8	68.5
DeiT-Tiny	1.3	72.2
PSViT-1D-Tiny	1.4	**77.4**
PSViT-2D-Tiny	1.3	**78.8**
ResNet50	4.1	76.1
ResNet50*	4.1	78.5
X50-32x4d	4.3	77.6
X50-32x4d*	4.3	79.1
DeiT-Small	4.6	79.6
PSViT-1D-Small	4.9	**80.7**
PSViT-2D-Small	4.4	**81.6**
X101-64x4d	15.6	79.6
X101-64x4d*	15.6	81.5
ViT-Base	17.6	77.9
DeiT-Base	17.6	81.8
PSViT-1D-Base	18.9	**82.6**
PSViT-2D-Base	15.5	**82.9**

• Experiments

PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- Attention Sharing, Token Pooling, AutoML에 대한 Ablation study
- ImageNet 데이터셋으로 실험

Attention Sharing	Token Pooling	AutoML	FLOPS (G)	acc-1D %	acc-2D %
			1.3	72.2	72.2
✓			1.3	73.8	-
	✓		1.3	76.3	76.7
✓	✓	✓	1.3	77.4	78.8

• Experiments

PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- Attention sharing 적용을 선택으로 두었기 때문에 정확한 효과를 평가하기 어려움
- 모든 층에 attention sharing을 무조건 적용하여 성능을 평가
- 인접한 두 층을 공유하는 것(sharing2)과 인접한 세 층을 공유하는 것(sharing3) 중 전자의 성능이 좋음
- 1층과 3층의 특징이 유의미하게 달라지기 때문에 1,2층을 공유하는 것보다 1~3층을 공유하는 것이 효과가 줄어듦

Table 5. ImageNet top-1 classification accuracy for different sharing settings. 'no sharing' denotes the network without sharing attention. 'sharing 2' and 'sharing 3' respectively denote every 2 and 3 adjacent transformer layers share the same attention map.

model	FLOPS (G)	acc-1D (%)
no sharing	1.3	72.2
sharing 2	1.3	73.8
sharing 3	1.3	73.1

• Experiments

PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- Attention Sharing, Token Pooling, AutoML에 대한 Ablation study
- ImageNet 데이터셋으로 실험

Attention Sharing	Token Pooling	AutoML	FLOPS (G)	acc-1D %	acc-2D %
			1.3	72.2	72.2
✓			1.3	73.8	-
	✓		1.3	76.3	76.7
✓	✓	✓	1.3	77.4	78.8

• Experiments

PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- 모델 사이즈(Tiny, Small) 및 패치 개수(8,16) 별 Auto ML을 통한 hyper-parameter 최적화 결과

Model	Tiny/8	Tiny/16	Small/8	Small/16
Token Dimensions	[64, 144, 192]	[192, 288, 384]	[144, 256, 384]	[288, 512, 768]
Num. heads	[1, 3, 3]	[3, 6, 6]	[3, 4, 6]	[6, 8, 12]
Token Numbers	[785, 393, 197]	[197, 99, 50]	[785, 393, 197]	[197, 99, 50]
Params(M)	3.8	15.6	13.9	53.8
FLOPS(G)	1.5	1.3	4.9	4.0
Top1 Acc	72.71%	77.40%	80.70 %	78.32 %

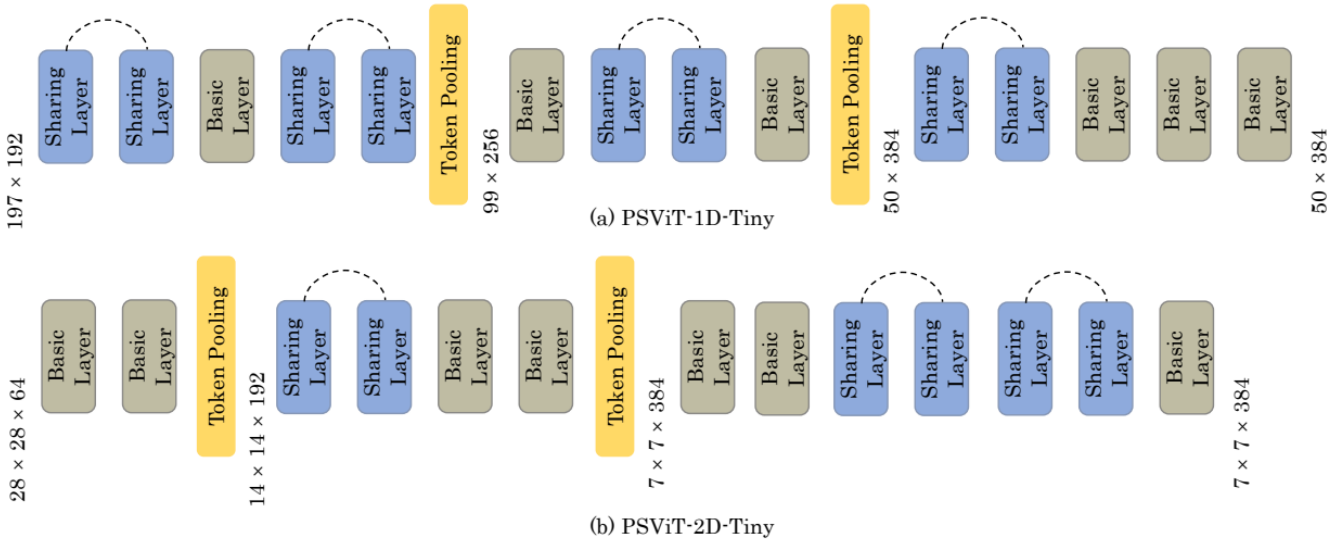


Figure 6. Searched architectures of PSViT-1D-Tiny and PSViT-2D-Tiny. The feature size does not change in the same stage.

• Experiments

PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- DeiT와의 GradCAM 비교

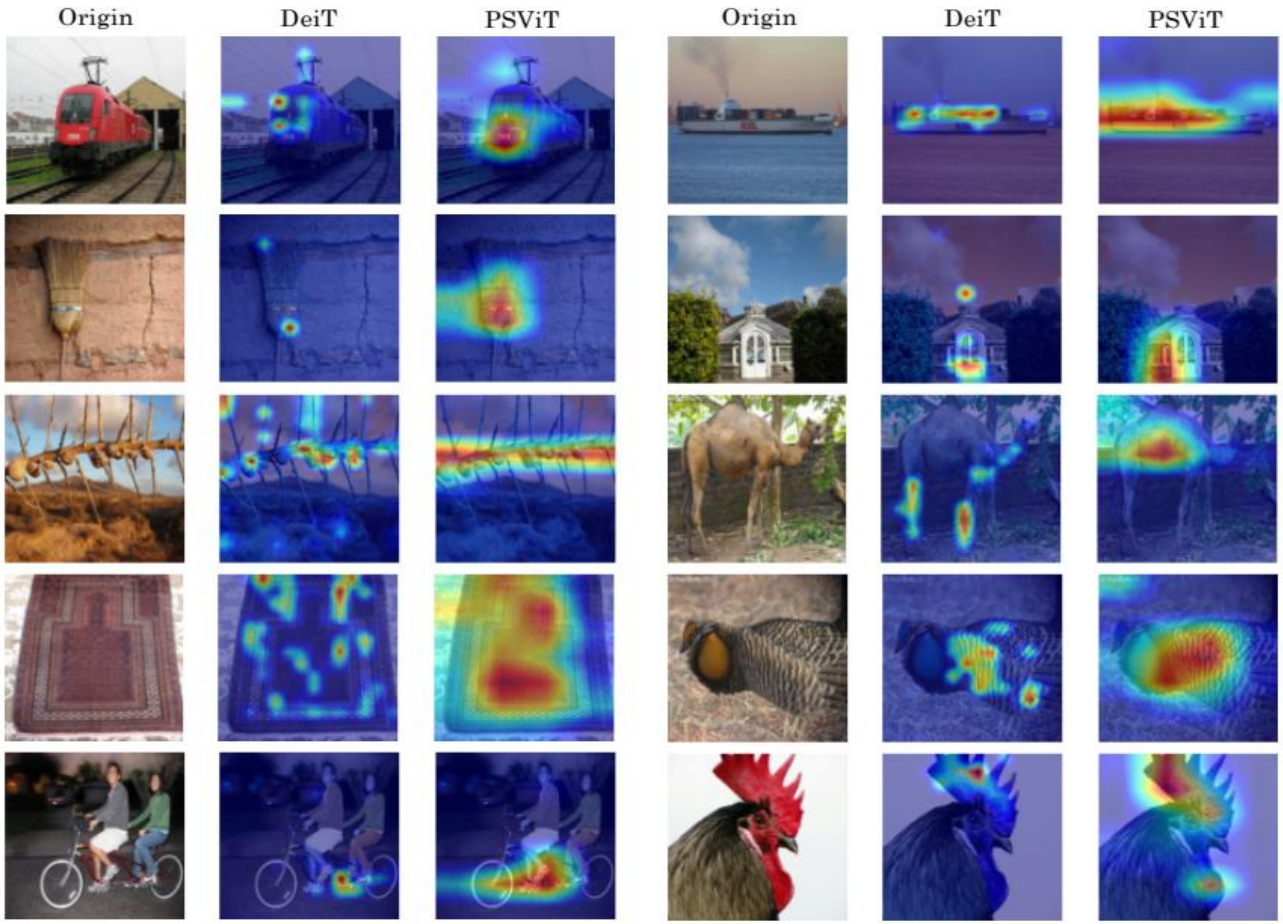


Figure 5. Visualization of features for DeiT and our PSViT. Images in the 1st and 4th columns are from ImageNet.

• Experiments

PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- Object detection and instance segmentation 실험 결과
- MSCOCO 2017 데이터셋과 Mask-RCNN-FPN framework를 사용

Table 6. Performance comparison between different backbones on the object detection and instance segmentation tasks.

Backbone	Object Detection						Instance Segmentation					
	AP	$AP50$	$AP75$	AP_s	AP_m	AP_l	AP	$AP50$	$AP75$	AP_s	AP_m	AP_l
ResNet18	34.8	56.3	37.5	20.1	37.0	45.2	32.8	53.4	34.8	17.4	25.1	44.7
PSViT-2D-Tiny	40.8	64.7	44.0	25.3	43.8	53.9	37.7	60.1	39.9	21.2	40.6	52.8
ResNet50	38.3	60.4	41.4	23.3	41.6	48.9	35.5	57.2	37.8	19.6	38.6	47.7
ResNet101	39.5	61.3	43.0	23.4	42.7	51.1	36.5	58.2	39.1	19.6	39.7	49.4

• Conclusion

❖ PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

- 트랜스포머가 중복된 특징을 뽑아내는 것을 발견하고 Token pooling과 Attention sharing이라는 간단한 기법을 적용하여 성능을 향상시킴
- 최적의 모델 구조를 찾기 위해서 AutoML을 적용

- 후기
  - ➢ Attention sharing의 효과를 입증했고 연산량을 줄여주지만 feature의 redundancy를 줄이는 것인지는 의문(같은 attention score를 공유하고 비슷한 embedding feature면 redundant한 특징이 추출되는 것이 아닌가?)
  - ➢ ViT가 점점 CNN과 닮아지고 있다는 생각이 들었음
  - ➢ Weight-based AutoML이 어떤 알고리즘으로 진행되는지 궁금

• Reference

PSViT : Better Vision Transformer via Token Pooling and Attention Sharing

1. Chen, B., Li, P., Li, B., Li, C., Bai, L., Lin, C., ... & Ouyang, W. (2021). PSViT: Better Vision Transformer via Token Pooling and Attention Sharing. *arXiv preprint arXiv:2108.03428*.
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

KOREA UNIVERSITY

DMQA