CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification

School of Industrial and Management Engineering, Korea University

Lee Kyung Yoo

KOREA UNIVERSITY

DMQA

Contents

Copyright ⓒ 2021, All rights reserved.

DMQA

Introduction

❖ CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification (arXiv, 2021)

- MIT-IBM Watson AI Lab에서 연구
- 2021년 10월 01일 기준으로 29회 인용

CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification

Chun-Fu (Richard) Chen, Quanfu Fan, Rameswar Panda
MIT-IBM Watson AI Lab
chenrich@us.ibm.com, qfan@us.ibm.com, rpanda@ibm.com

Abstract

The recently developed vision transformer (ViT) has achieved promising results on image classification compared to convolutional neural networks. Inspired by this, in this paper, we study how to learn multi-scale feature representations in transformer models for image classification. To this end, we propose a dual-branch transformer to combine image patches (i.e., tokens in a transformer) of different sizes to produce stronger image features. Our approach processes small-patch and large-patch tokens with two separate branches of different computational complexity and these tokens are then fused purely by attention multiple times to complement each other. Furthermore, to reduce computation, we develop a simple yet effective token fusion module based on cross attention, which uses a single token for each branch as a query to exchange information with other branches. Our proposed cross-attention only requires linear time for both computational and memory complexity instead of quadratic time otherwise. Extensive experiments demonstrate that our approach performs better than or on par with several concurrent works on vision transformer, in addition to efficient CNN models. For example, on the ImageNet1K dataset, with some architectural changes, our approach outperforms the recent DeiT by a large margin of 2% with a small to moderate increase in FLOPs and model parameters. Our source codes and models are available at https://github.com/IBM/CrossViT.

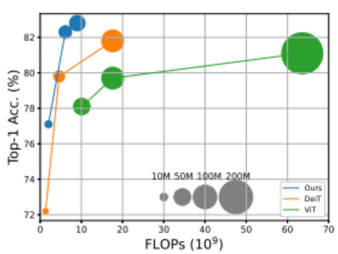


Figure 1: **Improvement of our proposed approach over DeiT [35] and ViT [11].** The circle size is proportional to the model size. All models are trained on ImageNet1K from scratch. The results of ViT are referenced from [45].

search efforts on transformers in vision have, until very recently, been largely focused on combining CNNs with self-attention [3, 48, 31, 32]. While these hybrid approaches achieve promising performance, they have limited scalability in computation compared to purely attention-based transformers. Vision Transformer (ViT) [11], which uses a sequence of embedded image patches as input to a standard transformer, is the first kind of convolution-free transform-

Copyright ⓒ 2021, All rights reserved.

DMQA

Introduction

❖ CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification (arXiv, 2021)

- Image classification 분야에서 multi-scale feature representation 학습을 위한 transformer 기반 모델 아키텍처를 새롭게 제안
- 기존 ViT구조를 기반으로, 다음과 같은 변형을 적용
  - ➢ **Multi-scale**을 활용한 **dual-branch transformer**
  - ➢ **Cross attention**을 활용한 **token fusion module**
- CNN 구조에서 긍정적 효과를 보인 multi-scale features를 ViT에 새롭게 적용하여 더 강력한 image feature 생성과 동시에 연산량 감소를 이루어냄

Copyright ⓒ 2021, All rights reserved.

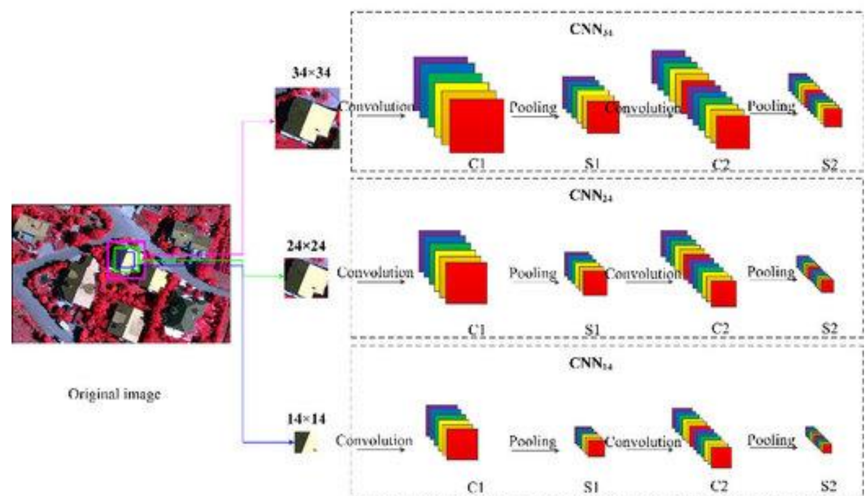
DMQA

Research Purpose

❖ Image classification을 위한 ViT 기반 multi-scale feature representation learning

- 다양한 스케일로 이미지의 특징을 추출하는 multi-scale feature representation은 주로 object detection 및 recognition 분야에서 주로 활용됨
- 또한 Big-Little Net, OctNet 등 일부 네트워크의 학습 속도를 빠르게 함
- 이러한 **multi-scale features representation을 image classification 분야에 적용**하는 것과 동시에 CNN 계열 모델이 아닌, 최근 우수한 성능을 보이고 있는 **ViT를 기반으로 한 새로운 모델 구조** 제안

CNN 계열
Multi-scale feature representation



Sun, Genyun, et al. "Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images." *Remote Sensing* 11.3 (2019): 227.

Copyright ⓒ 2021, All rights reserved.

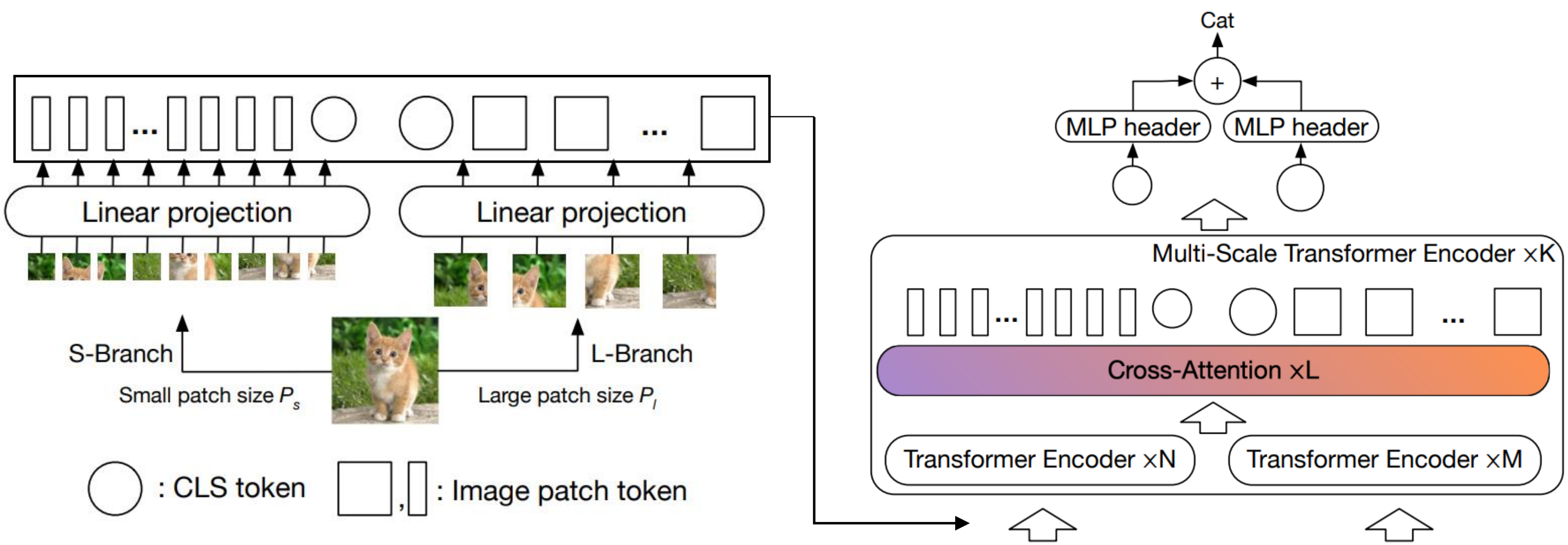
DMQA

CrossViT

- Overview of CrossViT

❖ Architecture

- K개의 multi-scale transformer encoders stack으로 구성
- 각각의 encoder는 두 개의 branch를 통해 서로 다른 크기의 image tokens을 처리하여 받음
- CLS tokens를 이용한 cross attention을 통해 최종적으로 각 image token을 융합함
- L번 융합되어 얻어진 CLS tokens를 최종 예측에 활용



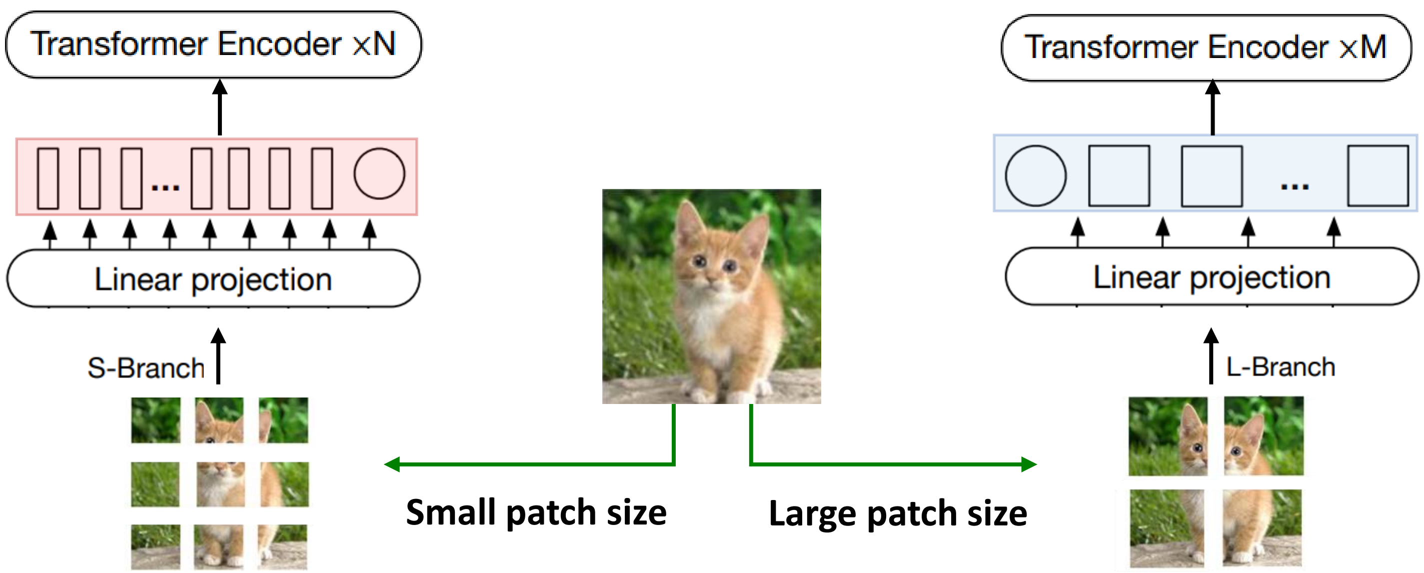
Copyright ⓒ 2021, All rights reserved.

CrossViT

- Two different branches

❖ Architecture

- S-Branch 상대적으로 적은 수의 encoder와 작은 embedding 차원으로 이루어진 fine-grained patch size (P_s) 로 linear projection layer 통과
- L-Branch: 상대적으로 많은 수의 encoder와 큰 embedding 차원으로 이루어진 coarse-grained patch size (P_l) 로 linear projection layer 통과
- 계산 비용의 균형을 맞추기 위해 두 branch에 다른 수(N, M)의 transformer encoder 적용



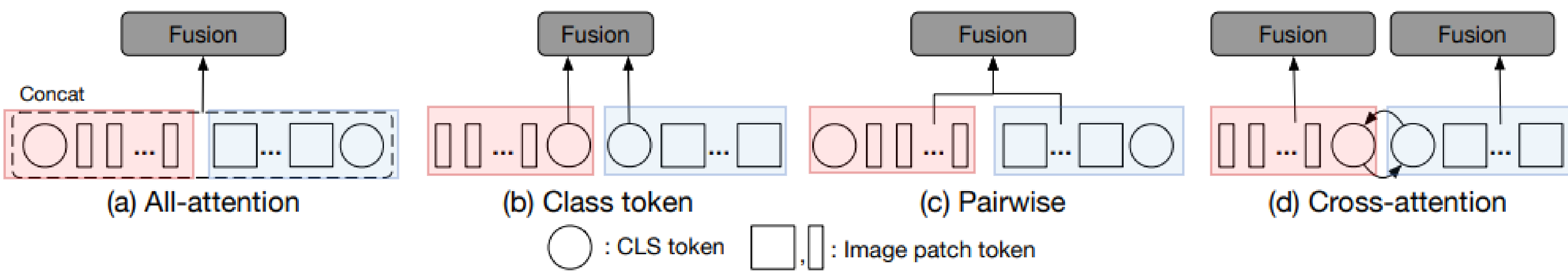
Copyright ⓒ 2021, All rights reserved.

CrossViT

- Multi-scale fusions

❖ Architecture

- 4가지의 서로 다른 fusion strategies를 비교하여 최적의 모듈로 cross-attention fusion 채택
- (a) All-attention fusion: 모든 토큰을 단순히 연결하여 self-attention module을 통해 융합
- (b) Class token fusion: Global feature representation으로 간주되는 CLS tokens만을 융합
- (c) Pairwise fusion: 이미지에서의 공간적 위치에 상응하는 토큰끼리 융합
- (d) Cross-attention fusion: 한 branch의 CLS token과 다른 branch의 patch token을 융합



Copyright ⓒ 2021, All rights reserved.

CrossViT

- Cross-attention module

❖ Architecture

- 각 branch의 CLS token을 정보를 담은 agent로 활용하여 다른 branch의 patch token과 결합하여 정보를 교환한 다음 이를 자체 branch로 가져와 backproject 진행
- 이를 통해 서로 다른 scale에서의 정보를 포함할 수 있으며, 다음 transformer encoder로 다시 입력됨에 따라 각각의 patch token의 representation이 풍부해짐

Cross-attention module for Large branch

$$\mathbf{x}'^{l} = \left[f^{l}(\mathbf{x}^{l}_{cls}) \,||\, \mathbf{x}^{s}_{patch}\right]$$

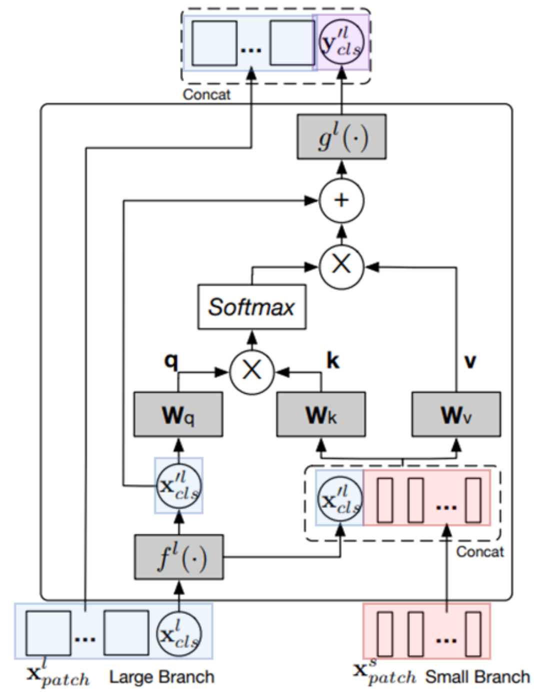
$$\mathbf{q} = \mathbf{x}'^{l}_{cls}\mathbf{W}_{q}, \quad \mathbf{k} = \mathbf{x}'^{l}\mathbf{W}_{k}, \quad \mathbf{v} = \mathbf{x}'^{l}\mathbf{W}_{v}$$

$$\mathbf{A} = \mathtt{softmax}(\mathbf{q}\mathbf{k}^{T}/\sqrt{C/h}), \quad \mathtt{CA}(\mathbf{x}'^{l}) = \mathbf{A}\mathbf{v}$$

$$\mathbf{y}^{l}_{cls} = f^{l}\left(\mathbf{x}^{l}_{cls}\right) + \mathtt{MCA}(\mathtt{LN}([f^{l}(\mathbf{x}^{l}_{cls}) \,||\, \mathbf{x}^{s}_{patch}]))$$

$$\mathbf{z}^{l} = \left[g^{l}\left(\mathbf{y}^{l}_{cls}\right) \,||\, \mathbf{x}^{l}_{patch}\right],$$

* f, g는 차원을 맞추기 위한 projection function



Experiments

❖ Architecture for model comparisons

- Image classification을 위한 여러 크기의 모델 구성
- 아래 하이퍼파라미터는 모든 종류의 모델에 관해 고정
- Multi-scale transformer encoder 수(K) = 3
- Small branch의 transformer encoder 수(N) = 1
- 한 multi-scale transformer encoder 내 존재하는 cross-attention module 수(L) = 1

Model	Patch embedding	Patch size Small	Patch size Large	Dimension Small	Dimension Large	# of heads	M	r
CrossViT-Ti	Linear	12	16	96	192	3	4	4
CrossViT-S	Linear	12	16	192	384	6	4	4
CrossViT-B	Linear	12	16	384	768	12	4	4
CrossViT-9	Linear	12	16	128	256	4	3	3
CrossViT-15	Linear	12	16	192	384	6	5	3
CrossViT-18	Linear	12	16	224	448	7	6	3
CrossViT-9†	3 Conv.	12	16	128	256	4	3	3
CrossViT-15†	3 Conv.	12	16	192	384	6	5	3
CrossViT-18†	3 Conv.	12	16	224	448	7	6	3

* M = Large branch의 transformer encoder 수

Copyright ⓒ 2021, All rights reserved.

DMQA

Experiments

❖ Default training settings

- 사용 데이터셋: ImageNet-1k(main) / CIFAR10,100, Pet, CropDisease, ChestXRay8(transfer)

	Main Results	Transfer
Batch size	4,096	768
Epochs	300	1,000
Optimizer	AdamW	SGD
Weight Decay	0.05	1e-4
Linear-rate Scheduler (Initial LR)	Cosine (0.004)	Cosine (0.01)
Warmup Epochs	30	5
Warmup linear-rate Scheduler (Initial LR)	Linear (1e-6)	
Data Aug.	RandAugment (m=9, n=2)	
Mixup (α)	0.8	
CutMix (α)	1.0	
Random Erasing	0.25	0.0
Instance Repetition*	3	
Drop-path	0.1	0.0
Label Smoothing	0.1	

*: only used for CrossViT-18.

Copyright ⓒ 2021, All rights reserved.

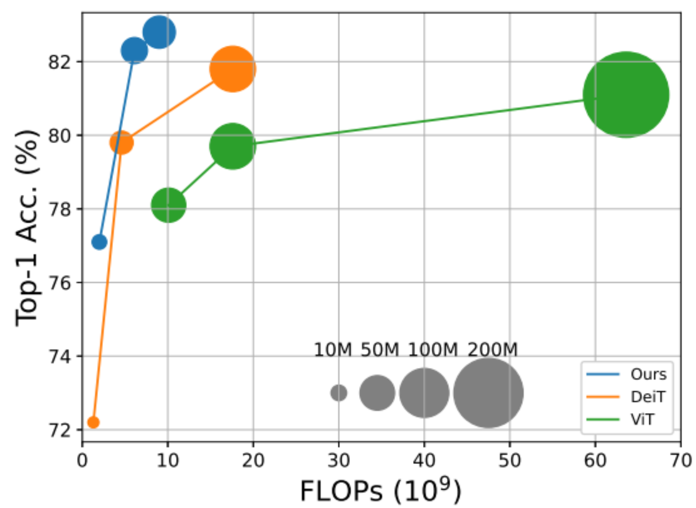
DMQA

Experiments

❖ Comparisons with DeiT baseline on ImageNet1K

- CrossViT가 상대적으로 적은 파라미터 수와 낮은 계산 복잡도로 Base model로 설정한 DeiT를 뛰어넘는 우수한 분류성능을 보여줌

Model	Top-1 Acc. (%)	FLOPs (G)	Throughput (images/s)	Params (M)
DeiT-Ti	72.2	1.3	2557	5.7
CrossViT-Ti	73.4 (+1.2)	1.6	1668	6.9
CrossViT-9	73.9 (+0.5)	1.8	1530	8.6
CrossViT-9†	**77.1** (+3.2)	2.0	1463	8.8
DeiT-S	79.8	4.6	966	22.1
CrossViT-S	81.0 (+1.2)	5.6	690	26.7
CrossViT-15	81.5 (+0.5)	5.8	640	27.4
CrossViT-15†	**82.3** (+0.8)	6.1	626	28.2
DeiT-B	81.8	17.6	314	86.6
CrossViT-B	82.2 (+0.4)	21.2	239	104.7
CrossViT-18	82.5 (+0.3)	9.0	430	43.3
CrossViT-18†	**82.8** (+0.3)	**9.5**	418	44.3



Copyright ⓒ 2021, All rights reserved.

DMQA

Experiments

❖ Comparisons with other recent transformer-based models on ImageNet1K

- DeiT를 제외한 타 transformer 계열의 모델과 비교하였을 때 역시 적은 파라미터 수와 낮은 계산 복잡도로 우수한 분류 성능을 보여줌

Model	Top-1 Acc. (%)	FLOPs (G)	Params (M)
Peceiver [19] (arXiv, 2021-03)	76.4	–	43.9
DeiT-S [35] (arXiv, 2020-12)	79.8	4.6	22.1
CentroidViT-S [42] (arXiv, 2021-02)	80.9	4.7	22.3
PVT-S [38] (arXiv, 2021-02)	79.8	3.8	24.5
PVT-M [38] (arXiv, 2021-02)	81.2	6.7	44.2
T2T-ViT-14 [45] (arXiv, 2021-01)	80.7	6.1*	21.5
TNT-S [14] (arXiv, 2021-02)	81.3	5.2	23.8
CrossViT-15 (Ours)	81.5	5.8	27.4
CrossViT-15† (Ours)	**82.3**	6.1	28.2
ViT-B@384 [11] (ICLR, 2021)	77.9	17.6	86.6
DeiT-B [35] (arXiv, 2020-12)	81.8	17.6	86.6
PVT-L [38] (arXiv, 2021-02)	81.7	9.8	61.4
T2T-ViT-19 [45] (arXiv, 2021-01)	81.4	9.8*	39.0
T2T-ViT-24 [45] (arXiv, 2021-01)	82.2	15.0*	64.1
TNT-B [14] (arXiv, 2021-02)	**82.8**	14.1	65.6
CrossViT-18 (Ours)	82.5	9.0	43.3
CrossViT-18† (Ours)	**82.8**	9.5	44.3

*: We recompute the flops by using our tools.

Copyright ⓒ 2021, All rights reserved.

Experiments

❖ Comparisons with CNN models on ImageNet1K

- CNN 계열의 모델과 비교하였을 때 역시 적은 파라미터 수와 낮은 계산 복잡도로 우수한 분류 성능을 보여줌

Model	Top-1 Acc. (%)	FLOPs (G)	Throughput (images/s)	Params (M)
ResNet-101 [15]	76.7	7.80	678	44.6
ResNet-152 [15]	77.0	11.5	445	60.2
ResNeXt-101-32×4d [43]	78.8	8.0	477	44.2
ResNeXt-101-64×4d [43]	79.6	15.5	289	83.5
SEResNet-101 [18]	77.6	7.8	564	49.3
SEResNet-152 [18]	78.4	11.5	392	66.8
SENet-154 [18]	81.3	20.7	201	115.1
ECA-Net101 [37]	78.7	7.4	591	42.5
ECA-Net152 [37]	78.9	10.9	428	59.1
RegNetY-8GF [30]	79.9	8.0	557	39.2
RegNetY-12GF [30]	80.3	12.1	439	51.8
RegNetY-16GF [30]	80.4	15.9	336	83.6
RegNetY-32GF [30]	81.0	32.3	208	145.0
EfficienetNet-B4@380 [34]	82.9	4.2	356	19
EfficienetNet-B5@456 [34]	83.7	9.9	169	30
EfficienetNet-B6@528 [34]	84.0	19.0	100	43
EfficienetNet-B7@600 [34]	84.3	37.0	55	66
CrossViT-15	81.5	5.8	640	27.4
CrossViT-15†	82.3	6.1	626	28.2
CrossViT-15†@384	83.5	21.4	158	28.5
CrossViT-18	82.5	9.03	430	43.3
CrossViT-18†	82.8	9.5	418	44.3
CrossViT-18†@384	83.9	32.4	112	44.6
CrossViT-18†@480	84.1	56.6	57	44.9

Copyright ⓒ 2021, All rights reserved.

DMQA

Experiments

❖ Transfer learning performance

- CrossViT 은 검증한 모든 downstream tasks에서 최신 DeiT 모델의 분류 성능에 준하는 경쟁력 있는 좋은 결과를 보임

Model	CIFAR10	CIFAR100	Pet	CropDiseases	ChestXRay8
DeiT-S [35]	99.15	90.89	94.93	99.96	55.39
DeiT-B [35]	99.10*	90.80*	94.39	99.96	55.77
CrossViT-15	99.00	90.77	94.55	99.97	55.89
CrossViT-18	99.11	91.36	95.07	99.97	55.94

*: numbers reported in the original paper.

Copyright ⓒ 2021, All rights reserved.

DMQA

Conclusion

❖ Conclusion

- Image의 분할 정도는 image classification task에서 ViT의 정확도와 복잡도에 주요한 영향을 미침
- CrossViT은 서로 다른 scale의 image patch token을 효과적으로 잘 결합하여 두 branch의 서로 다른 feature representation에 대한 정보를 교환할 수 있게 함
- 이를 통해 CrossViT는 우수한 accuracy를 보이면서도, 낮은 computational cost를 유지함
- Multi-scale feature representation을 image classification 분야에서 ViT에 새롭게 적용한 구조가 타 논문에 비해 신선했음

Copyright ⓒ 2021, All rights reserved.

DMQA

References

- Chen, Chun-Fu, et al. "Big-little net: An efficient multi-scale feature representation for visual and speech recognition." arXiv preprint arXiv:1807.03848 (2018).
- Riegler, Gernot, Ali Osman Ulusoy, and Andreas Geiger. "Octnet: Learning deep 3d representations at high resolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- Sun, Genyun, et al. "Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images." Remote Sensing 11.3 (2019): 227.

Copyright ⓒ 2021, All rights reserved.

Thank You