Focal Self-attention for Local-Global Interactions in Vision Transformers

School of Industrial and Management Engineering, Korea University

Jae Hoon Kim

KOREA UNIVERSITY

DMQA

Contents

- Research Purpose
- Focal Transformers
- Experiments
- Conclusion

Research Purpose

❖ Focal Self-attention for Local-Global Interactions in Vision Transformers (arXiv, 2021)

- Microsoft에서 연구하였으며 2021년 09월 19일 기준으로 2회 인용됨

Focal Self-attention for Local-Global Interactions in Vision Transformers

Jianwei Yang[1] **Chunyuan Li**[1] **Pengchuan Zhang**[1] **Xiyang Dai**[2] **Bin Xiao**[2]
Lu Yuan[2] **Jianfeng Gao**[1]
[1]Microsoft Research at Redmond, [2]Microsoft Cloud + AI
{jianwyan,chunyl,penzhan,xidai,bixi,luyuan,jfgao}@microsoft.com

Abstract

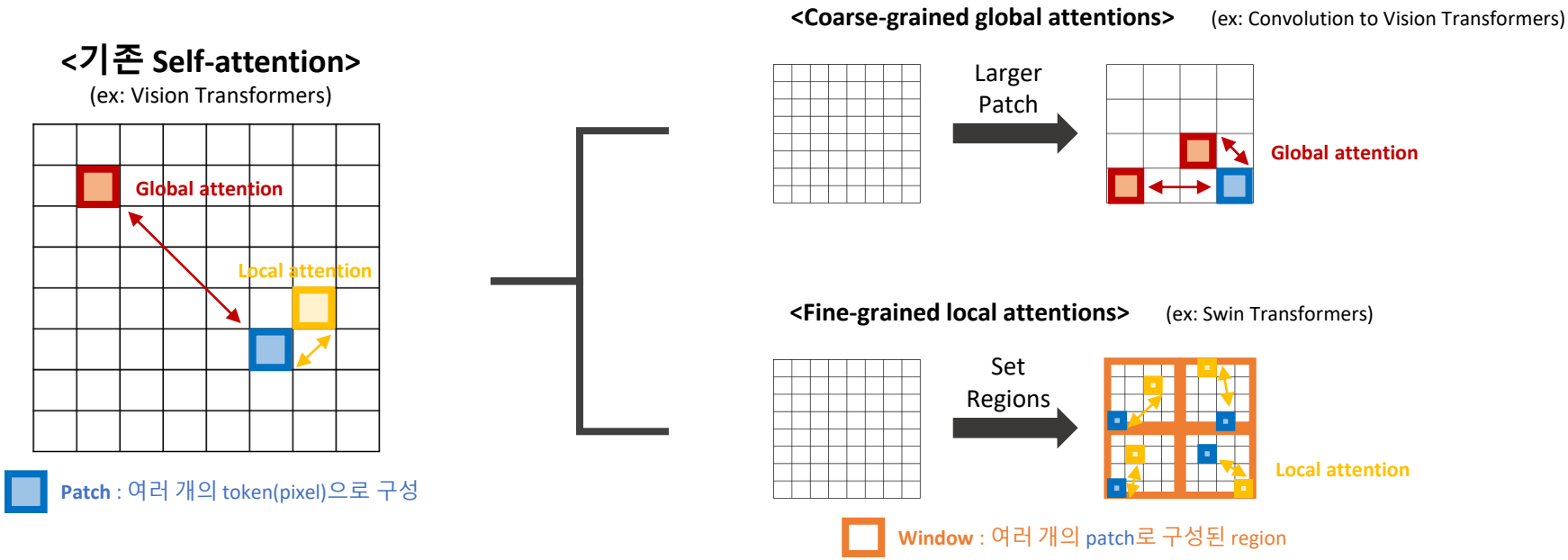
Recently, Vision Transformer and its variants have shown great promise on various computer vision tasks. The ability of capturing short- and long-range visual dependencies through self-attention is the key to success. But it also brings challenges due to quadratic computational overhead, especially for the high-resolution vision tasks (*e.g.*, object detection). Many recent works have attempted to reduce the computational and memory cost *and* improve performance by applying either coarse-grained global attentions or fine-grained local attentions. However, both approaches cripple the modeling power of the original self-attention mechanism of multi-layer Transformers, thus leading to sub-optimal solutions. In this paper, we present *focal self-attention*, a new mechanism that incorporates both fine-grained local and coarse-grained global interactions. In this new mechanism, each token attends its closest surrounding tokens at fine granularity and the tokens far away at coarse granularity, and thus can capture both short- and long-range visual dependencies efficiently *and* effectively. With focal self-attention, we propose a new variant of Vision Transformer models, called *Focal Transformer*, which achieves superior performance over the state-of-the-art (SoTA) vision Transformers on a range of public image classification and object detection benchmarks. In particular, our Focal Transformer models with a moderate size of 51.1M and a larger size of 89.8M achieve 83.5% and 83.8% Top-1 accuracy, respectively, on ImageNet classification at 224×224. When employed as the backbones, Focal Transformers achieve consistent and substantial improvements over the current SoTA Swin Transformers [44] across 6 different object detection methods. Our largest Focal Transformer yields **58.7/58.9** box mAPs and **50.9/51.3** mask mAPs on COCO mini-val/test-dev, and **55.4** mIoU on ADE20K for semantic segmentation, creating new SoTA on three of the most challenging computer vision tasks.

DMQA

Research Purpose

❖ Focal Self-attention for Local-Global Interactions in Vision Transformers (arXiv, 2021)

- 기존 self-attention은 연산량 문제(quadratic computation cost)로 인해 고해상도 이미지 처리가 힘듦
- 이전에는 기존 self-attention의 연산량 문제를 local 혹은 global self-attention 중 하나에 집중하여 해결
- 하지만 local 또는 global self-attention만을 사용하면 Transfomers 구조에서 sub-optimal한 성능을 냄
- 따라서 optimal한 성능을 낼 수 있도록 기존 self-attention을 연산 효율적으로 수행하는 구조를 제안

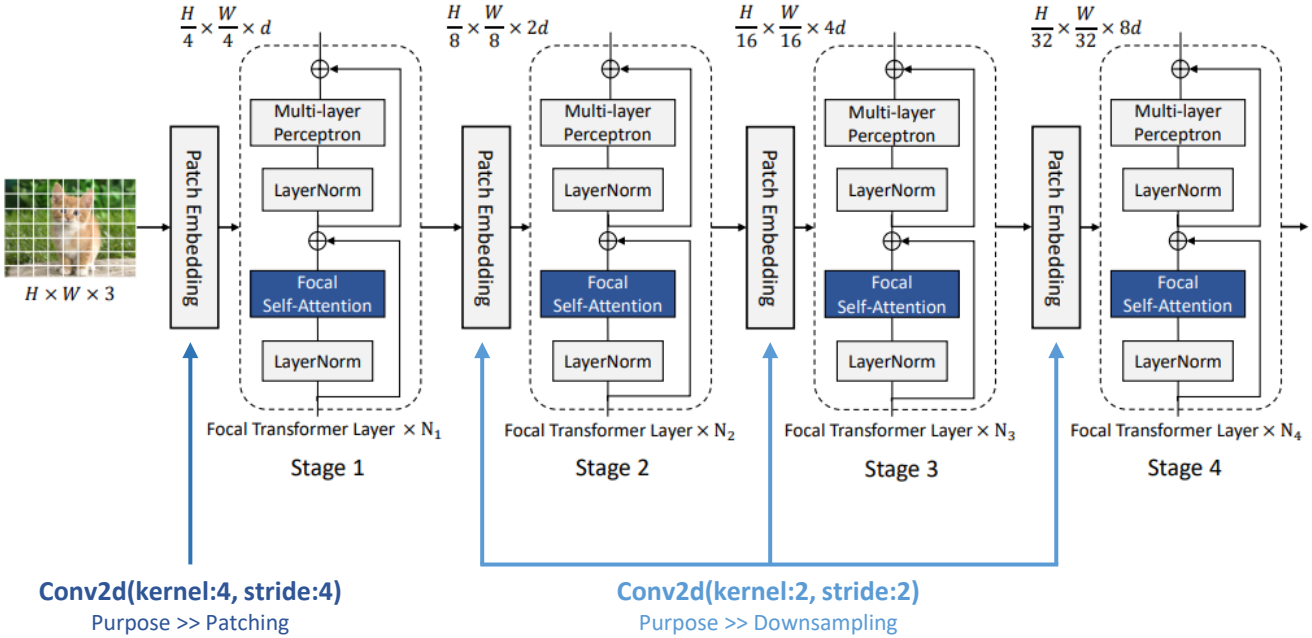


DMQA

Focal Transformers

❖ Diagram of Focal Transformers (overall architectures)

- 논문에서는 총 4개의 stage로 구성함
- Patch embedding은 convolution layer로 구성되며 input에 적용될 때와 stage 사이에 적용될 때 서로 다른 하이퍼 파라미터를 가짐

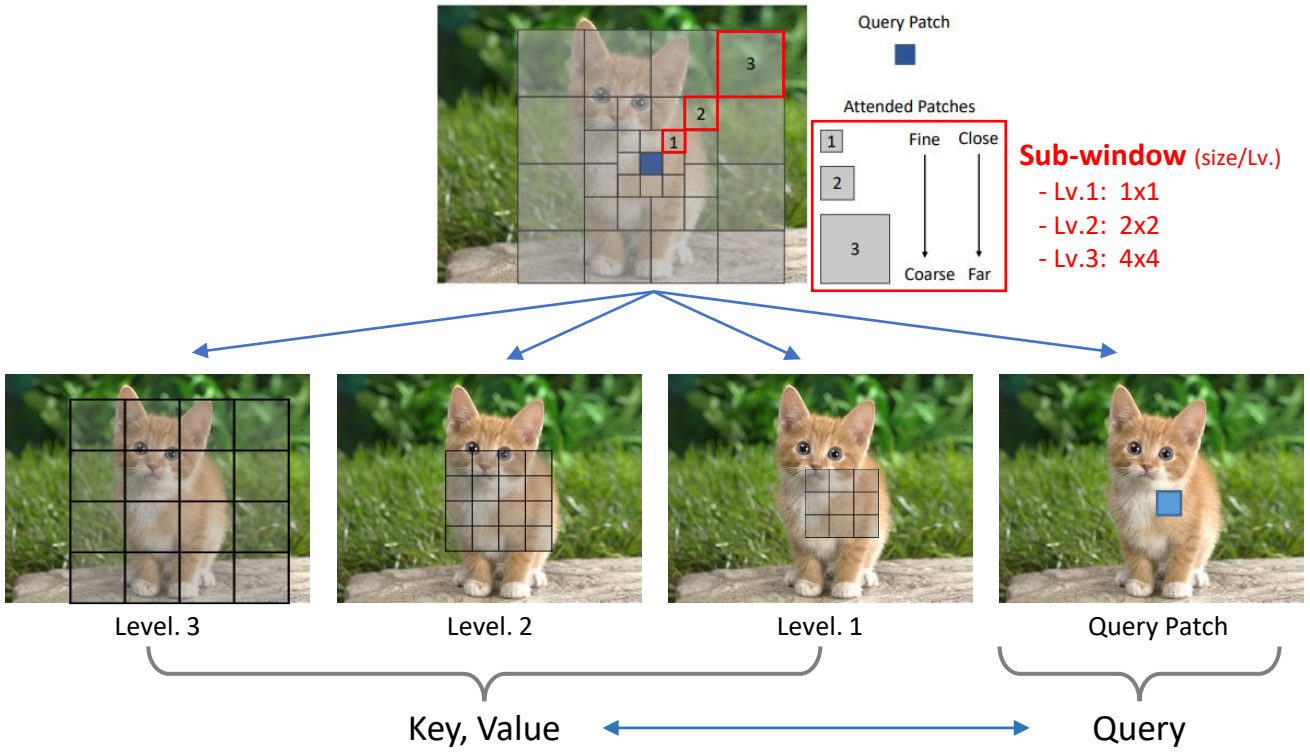


DMQA

Focal Transformers

❖ Diagram of Focal Transformers (window-wise focal self-attention)

- Query 역할을 할 patch를 지정한 뒤 이를 중심으로 둘러싸는 다양한 크기의 window region 지정
- Sub-window로부터 생성된 key, value와 지정한 query patch를 통해서 attention 연산을 수행

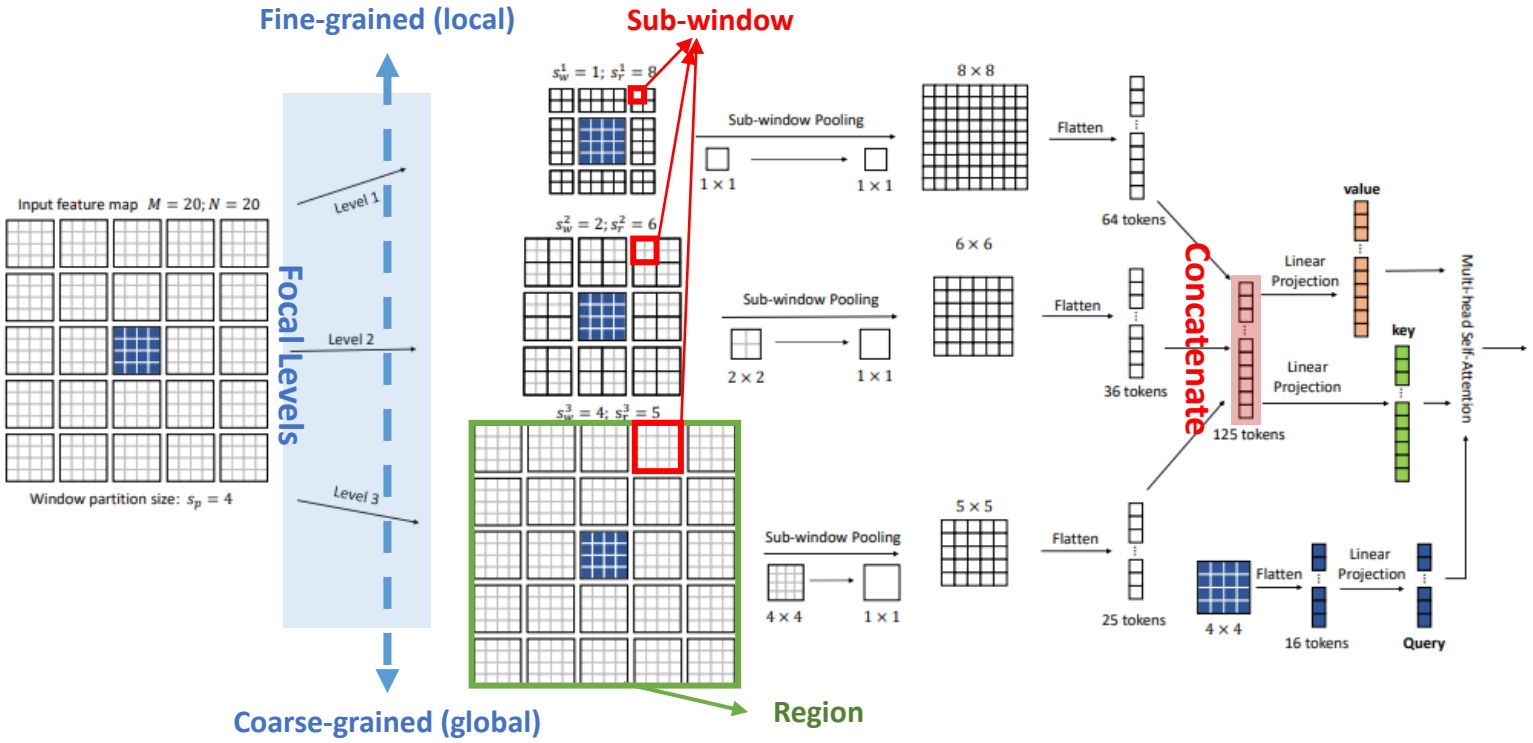


DMQA

Focal Transformers

❖ Diagram of Focal Transformers (window-wise focal self-attention)

- Focal **levels** (L) : Query patch 주변의 정보를 추출할 입도(fine 혹은 coarse)의 단계
- Focal **window** size (s_w^l) : Sub-window의 크기 → n x n개의 patch로 구성됨
- Focal **region** size (s_r^l) : Region의 크기 → m x m 개의 sub-window로 구성됨

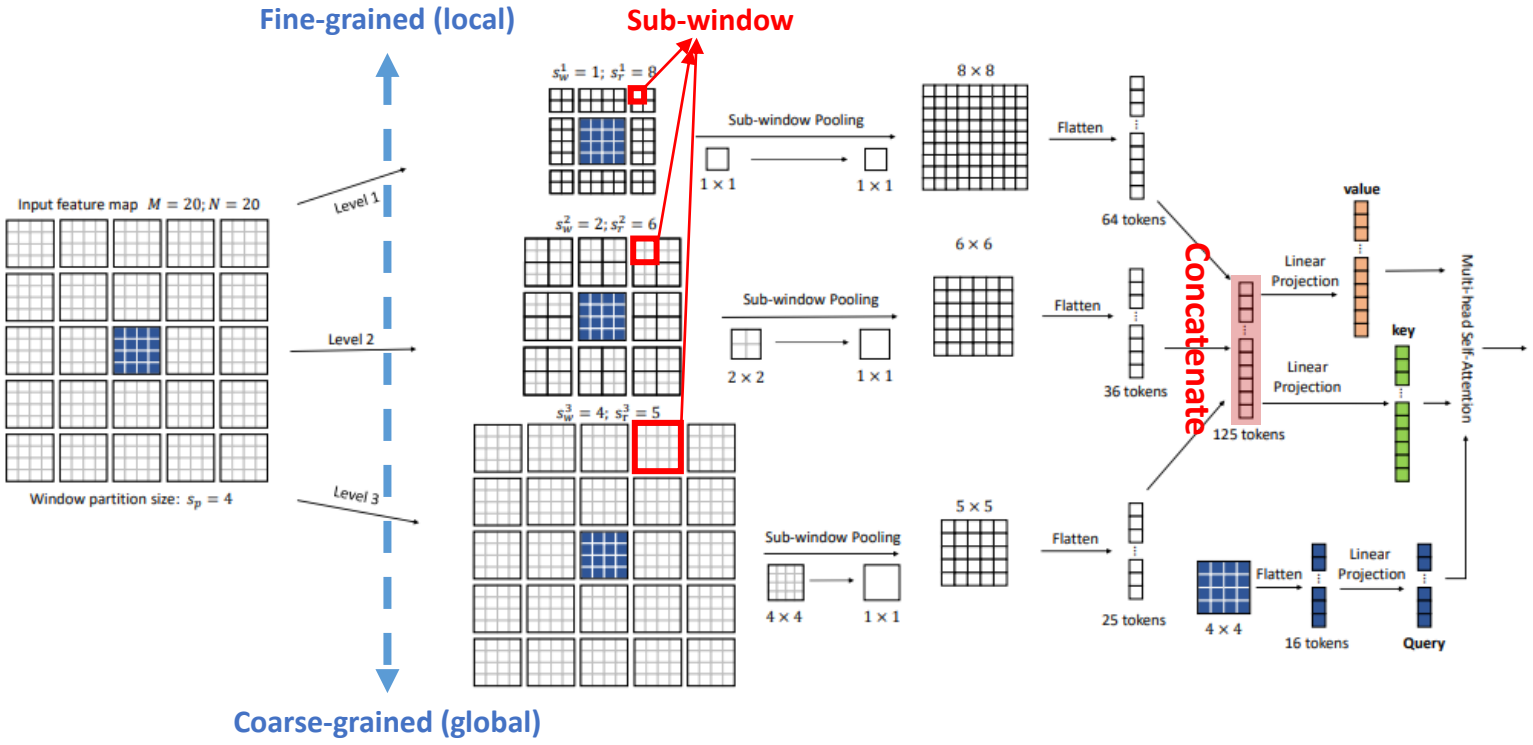


DMQA

Focal Transformers

❖ Diagram of Focal Transformers (window-wise focal self-attention)

- Sub-window pooling을 통해 window 단위로 feature map의 정보를 요약함
- 각 level 별 요약된 정보를 결합하여 local 및 global 정보를 모두 갖춘 representation vector를 생성



DMQA

Focal Transformers

❖ Diagram of Focal Transformers (window-wise focal self-attention)

- Focal attention은 query patch를 둘러싸는 크기의 window에 대하여 global attention을 계산함
- 반면 self-attention은 query patch와 동일한 크기의 patch에 대하여 attention을 계산함
- Query patch를 구성하는 token의 개수가 늘어날수록 focal attention의 global attention을 계산하는 window의 크기 또한 커짐
- 따라서 token의 개수가 증가할수록 focal attention의 receptive field가 더 크게 증가할 수 있음

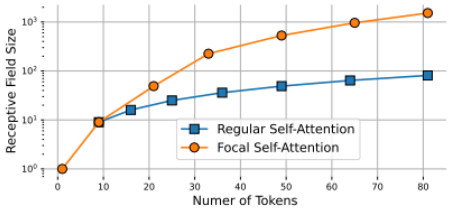
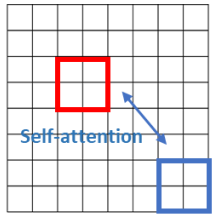
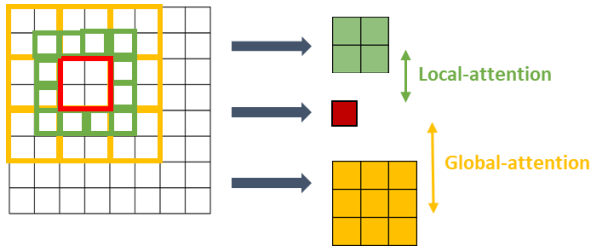


Figure 3: The size of receptive field (y-axis) with the increase of used tokens (x-axis) for standard and our focal self-attention. For focal self-attention, we assume increasing the window granularity by factor 2 gradually but no more than 8. Note that the y-axis is logarithmic.





DMQA

Experiments

❖ Image classification on ImageNet-1K

- 해당 실험에서 주목할 부분은 Focal-Small 성능과 다른 모델의 Base 성능과의 비교임
- Focal Transformer가 보다 적은 파라미터 수와 낮은 계산 복잡도로 월등한 성능을 보여줌

Model	#Params.	FLOPs	Top-1 (%)
ResNet-50 [34]	25.0	4.1	76.2
DeiT-Small/16 [57]	22.1	4.6	79.9
PVT-Small [63]	24.5	3.8	79.8
ViL-Small [80]	24.6	5.1	82.0
CvT-13 [67]	20.0	4.5	81.6
Swin-Tiny [44]	28.3	4.5	81.2
Focal-Tiny (Ours)	29.1	4.9	82.2
ResNet-101 [34]	45.0	7.9	77.4
PVT-Medium [63]	44.2	6.7	81.2
CvT-21 [67]	32.0	7.1	82.5
ViL-Medium [80]	39.7	9.1	83.3
Swin-Small [44]	49.6	8.7	83.1
Focal-Small (Ours)	51.1	9.1	83.5
ResNet-152 [34]	60.0	11.0	78.3
ViT-Base/16 [23]	86.6	17.6	77.9
DeiT-Base/16 [57]	86.6	17.5	81.8
PVT-Large [63]	61.4	9.8	81.7
ViL-Base [80]	55.7	13.4	83.2
Swin-Base [44]	87.8	15.4	83.4
Focal-Base (Ours)	89.8	16.0	**83.8**

Table 2: Comparison of image classification on ImageNet-1K for different models. Except for ViT-Base/16, all other models are trained and evaluated on 224×224 resolution.

Default training settings (all models)

- Dataset: ImageNet-1k
- Optimizer: AdamW (weight decay: 0.05)
- Batch size: 1024
- Epoch: 300
- Initial learning rate: 0.001
- Linear warm-up rate: 0.00001 (20 epoch)
- Cosine learning scheduler
- Stochastic depth drop rates: [0.2, 0.2, 0.3] each for tiny, small, base model
- Augmentation: training: random crop / valid, test: center crop

Experiments

❖ Object detection and instance segmentation (COCO)

- 해당 task에서는 focal size를 (15, 13, 9, 7)로 증가하여 진행함
  - ✓ Stage 1, 2에서는 focal attention의 범위가 이미지의 절반 이상이 되도록 보장되며,
  - ✓ Stage 3, 4 에서는 focal attention의 범위가 이미지 전체가 되도록 보장됨
- Classification task와 마찬가지로 Focal-Small이 Base 모델보다 적은 파라미터 수와 낮은 계산 복잡도를 가지고 더 좋은 성능을 보임

Backbone	#Params (M)	FLOPs (G)	RetinaNet 3x schedule + MS AP^b	AP^b_{50}	AP^b_{75}	AP_S	AP_M	AP_L	Mask R-CNN 3x schedule + MS AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet50 [34]	37.7/44.2	239/260	39.0	58.4	41.8	22.4	42.8	51.6	41.0	61.7	44.9	37.1	58.4	40.1
PVT-Small[63]	34.2/44.1	226/245	42.2	62.7	45.0	26.2	45.2	57.2	43.0	65.3	46.9	39.9	62.5	42.8
ViL-Small [80]	35.7/45.0	252/174	42.9	63.8	45.6	27.8	46.4	56.3	43.4	64.9	47.0	39.6	62.1	42.4
Swin-Tiny [44]	38.5/47.8	245/264	45.0	65.9	48.4	29.7	48.9	58.1	46.0	68.1	50.3	41.6	65.1	44.9
Focal-Tiny (Ours)	39.4/48.8	265/291	**45.5**	**66.3**	**48.8**	**31.2**	**49.2**	**58.7**	**47.2**	**69.4**	**51.9**	**42.7**	**66.5**	**45.9**
ResNet101 [34]	56.7/63.2	315/336	40.9	60.1	44.0	23.7	45.0	53.8	42.8	63.2	47.1	38.5	60.1	41.3
ResNeXt101-32x4d [70]	56.4/62.8	319/340	41.4	61.0	44.3	23.9	45.5	53.7	44.0	64.4	48.0	39.2	61.4	41.9
PVT-Medium [63]	53.9/63.9	283/302	43.2	63.8	46.1	27.3	46.3	58.9	44.2	66.0	48.2	40.5	63.1	43.5
ViL-Medium [80]	50.8/60.1	339/261	43.7	64.6	46.4	27.9	47.1	56.9	44.6	66.3	48.5	40.7	63.8	43.7
Swin-Small [44]	59.8/69.1	335/354	46.4	67.0	50.1	31.0	50.1	60.3	48.5	70.2	53.5	43.3	67.3	46.6
Focal-Small (Ours)	61.7/71.2	367/401	**47.3**	**67.8**	**51.0**	**31.6**	**50.9**	**61.1**	**48.8**	**70.5**	**53.6**	**43.8**	**67.7**	**47.2**
ResNeXt101-64x4d [70]	95.5/102	473/493	41.8	61.5	44.4	25.2	45.4	54.6	44.4	64.9	48.8	39.7	61.9	42.6
PVT-Large[63]	71.1/81.0	345/364	43.4	63.6	46.1	26.1	46.0	59.5	44.5	66.0	48.3	40.7	63.4	43.7
ViL-Base [80]	66.7/76.1	443/365	44.7	65.5	47.6	29.9	48.0	58.1	45.7	67.2	49.9	41.3	64.4	44.5
Swin-Base [44]	98.4/107	477/496	45.8	66.4	49.1	29.9	49.4	60.3	48.5	69.8	53.2	43.4	66.8	46.9
Focal-Base (Ours)	100.8/110.0	514/533	**46.9**	**67.8**	**50.3**	**31.9**	**50.3**	**61.5**	**49.0**	**70.1**	**53.6**	**43.7**	**67.6**	**47.0**

Table 4: COCO object detection and segmentation results with RetinaNet [42] and Mask R-CNN [34]. All models are trained with 3× schedule and multi-scale inputs (MS). The numbers before and after "/" at column 2 and 3 are the model size and complexity for RetinaNet and Mask R-CNN, respectively.

Default training settings (all models)

- Dataset: COCO 2017
- Optimizer: AdamW (weight decay: 0.05)
- Epoch: 36 (12 epoch per 1x schedule)
- Initial learning rate: 0.0001
- Stochastic depth drop rates: [0.2, 0.2, 0.3] each for tiny, small, base model
- Augmentation: resize short part [480 ~ 800], long part [~ 1333]

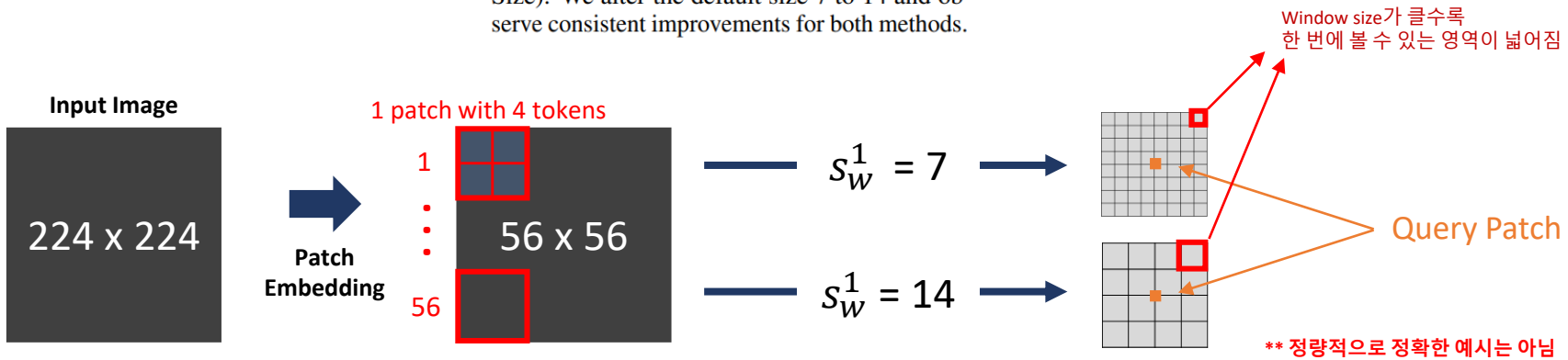
Experiments

❖ Ablation studies (Effect of varying window size)

- Window 너비를 늘려 receptive field를 넓히면 모델의 성능이 올라갈 것인가에 대한 실험
- Window 내에 들어오는 patch의 개수가 많아지면 query patch와의 attention을 계산할 때 더 넓은 범위(receptive field)에 대한 long-range interaction (global attention)이 이뤄짐
- 따라서 실험 결과를 볼 때, long-range interaction을 넓혀서 성능 향상을 꾀할 수 있음

Model	W-Size	FLOPs	Top-1 (%)	AP^b	AP^m
Swin-Tiny	7	4.5	81.2	43.7	39.8
	14	4.9	82.1	44.0	40.5
Focal-Tiny	7	4.9	82.2	44.9	41.1
	14	5.2	82.3	45.5	41.5

Table 8: Impact of different window sizes (W-Size). We alter the default size 7 to 14 and observe consistent improvements for both methods.



DMQA

Experiments

❖ Ablation studies (The necessity of window shift)

- Swin transformer는 window가 고정되는 한계를 극복하기 위해서 window shifting을 사용함
- Focal transformer는 query patch를 기준으로 fine, coarse grain window가 계속해서 달라짐 (고정 X)
- 실험 결과 또한 뒷받침 해주 듯 window shifting의 유무는 focal transformer에 별 영향을 주지 않음

Model	W-Shift	Top-1 (%)	AP^b	AP^m
Swin-Tiny	-	80.2	38.8	36.4
	✓	81.2	43.7	39.8
Focal-Tiny	-	82.2	44.8	41.0
	✓	81.9	44.9	41.1

Table 9: Impact of window shift (W-Shift) on Swin Transformer and Focal Transformer. Tiny models are used.

Experiments

❖ Ablation studies (Contributions of short- and long-range interaction)

- 단순히 window만 설정할 경우에는 고정된 영역만을 분석함
- 따라서 입력 이미지의 context 정보를 잡아내지 못하기 때문에 성능이 낮아짐

- 한편 object detection의 경우 global context보다 local context에 더 의존하는 경향이 있음
- 따라서 local attention을 추가했을 때보다 global attention을 추가했을 때 성능 향상의 폭이 더 큼

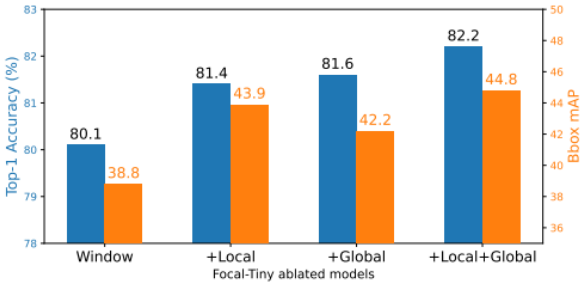


Figure 5: Ablating Focal-Tiny model by adding local, global and both interactions, respectively. Blue bars are for image classification and orange bars indicate object detection performance. Both local and global interactions are essential to obtain good performance. Better viewed in color.

Experiments

❖ Ablation studies (Model capacity against model depth)

- Swin transformer는 global attention을 수행하지 않음
- Global attention을 수행하지 않는 모델에 비해서 focal transformer 구조의 효율/효과를 비교 실험
- Third stage에서 focal transformer의 depth가 2 낮음에도 swin transformer보다 비슷하거나 높게 나옴

Depths	Model	#Params.	FLOPs	Top-1 (%)	AP^b	AP^m
2-2-2-2	Swin	21.2	3.1	78.7	38.2	35.7
	Focal	21.7	3.4	79.9	40.5	37.6
2-2-4-2	Swin	24.7	3.8	80.2	41.2	38.1
	Focal	25.4	4.1	81.4	43.3	39.8
2-2-6-2	Swin	28.3	4.5	81.2	43.7	39.8
	Focal	29.1	4.9	82.2	44.8	41.0

Table 10: Impact of the change of model depth. We gradually reduce the number of transformer layers at the third stage from original 6 to 4 and further 2. It apparently hurts the performance but our Focal Transformers has much slower drop rate than Swin Transformer.

Conclusion

❖ Conclusion & Limitations

- Local-global interaction을 효율적으로, 효과적으로 수행할 수 있는 focal self-attention을 제안함

- 하지만 본 모델의 밑바탕이 되는 swin transformer에 비해 같은 크기의 모델에서 더 많은 연산량과 메모리를 요구한다는 단점이 있음 (coarse-grained global attention이 추가 되었기 때문)

- 또한 query patch 주변의 sub-window로부터 feature map을 추출하는 과정으로 인해 swin transformer보다 학습 시간이 오래 걸린다는 단점이 있음

Reference

1. Yang, Jianwei, et al. "Focal self-attention for local-global interactions in vision transformers." *arXiv preprint arXiv:2107.00641* (2021).


Thank You