MetaFormer Is Actually What You Need for Vision

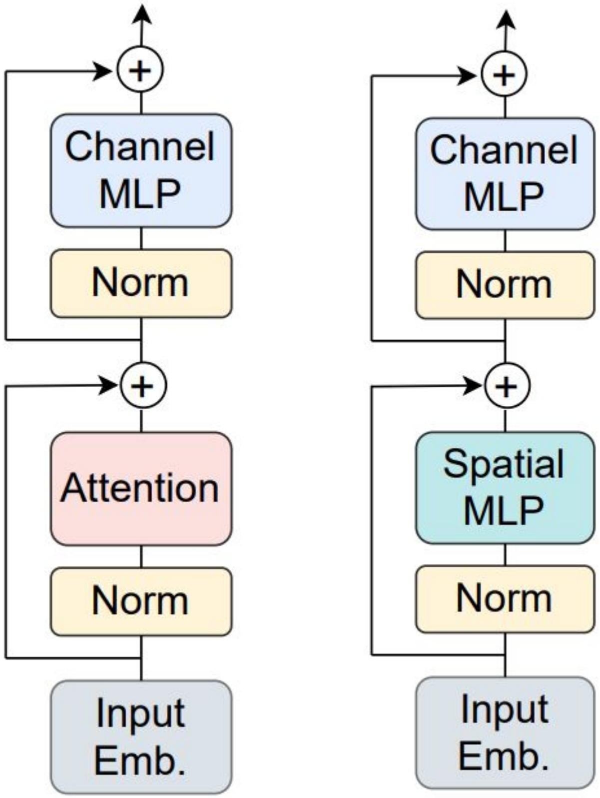
Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou,
Xinchao Wang, Jiashi Feng, and Shuicheng Yan

CVPR 2022

Presented by: Nicholas and Tarik

Background

- Previous ViT architecture uses attention and channel MLP layers with normalization
- The self-attention is quadratic to the number of tokens
- Another work replaced the self-attention with spatial MLP to achieve competitive results
- This step is referred to as the token mixer

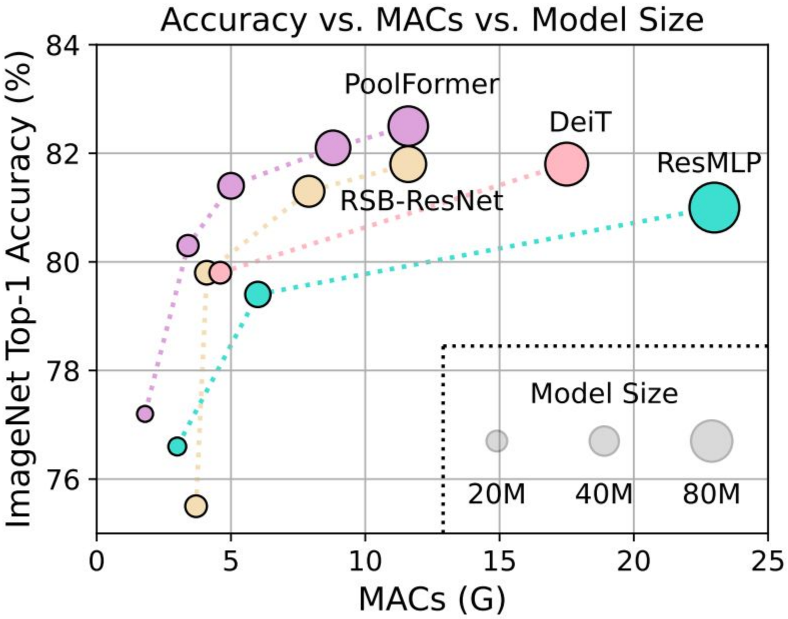


Transformer
(e.g. DeiT)

MLP-like model
(e.g. ResMLP)

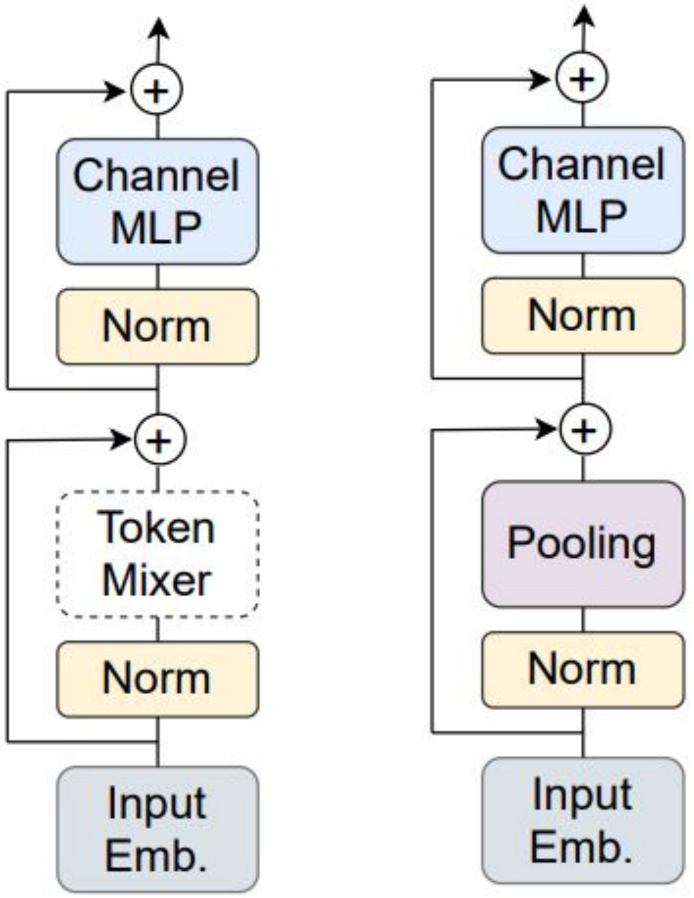
Motivation

- Success of transformers are often attributed to the self-attention token mixer
- However, the MetaFormer architecture is what is required for the competitive performance
- Using a very simple pooling token mixer, PoolFormer outperforms other models



Motivation

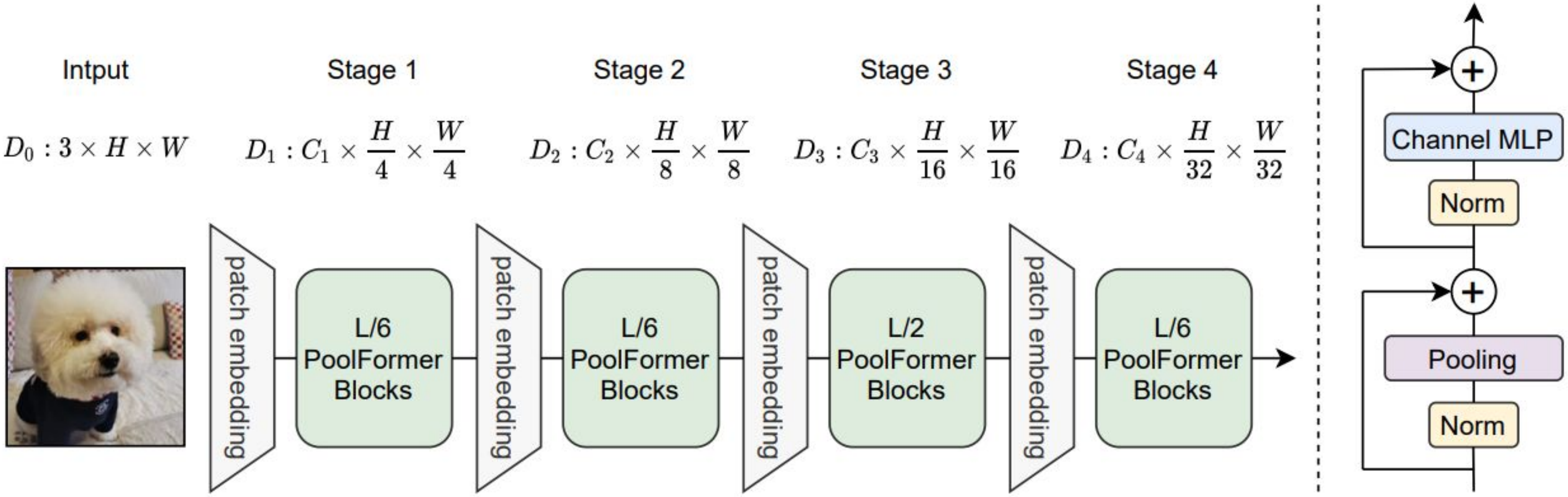
- The MetaFormer architecture is a generalized form of a transformer
- The PoolFormer is a specific instance of the MetaFormer that uses simple pooling as the token mixer



MetaFormer (General Arch.)   **PoolFormer** (Ours)

Methodology: Overview

- The pooling is made of 4 stages, with L total pooling blocks
- Pooling is a linear computational complexity algorithm



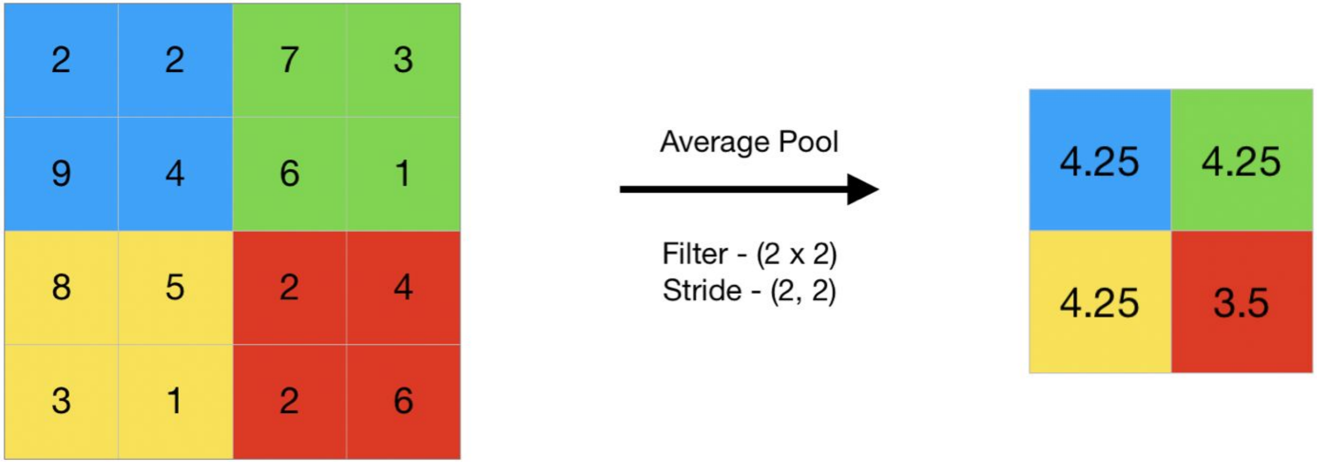
(a) Overal framework with L PoolFormer blocks

(b) PoolFormer block

Methodology: What is pooling?

- The formula for pooling with 3D data ($T_{:,i,j}$) is:

$$T'_{:,i,j} = \frac{1}{K \times K} \sum_{p,q=1}^{K} T_{:,i+p-\frac{K+1}{2},i+q-\frac{K+1}{2}} - T_{:,i,j}, \quad (4)$$



Methodology

- Multiple PoolFormer models are trained
- The hyperparameters are listed in the table
- Named “S” and “M” for small and medium embedding sizes
- L is the the number of pooling blocks

Stage	#Tokens	Layer Specification		PoolFormer				
				S12	S24	S36	M36	M48
1	$\frac{H}{4} \times \frac{W}{4}$	Patch Embedding	Patch Size	7×7, stride 4				
			Embed. Dim.	64			96	
		PoolFormer Block	Pooling Size	3×3, stride 1				
			MLP Ratio	4				
			# Block	2	4	6	6	8
2	$\frac{H}{8} \times \frac{W}{8}$	Patch Embedding	Patch Size	3×3, stride 2				
			Embed. Dim.	128			192	
		PoolFormer Block	Pooling Size	3×3, stride 1				
			MLP Ratio	4				
			# Block	2	4	6	6	8
3	$\frac{H}{16} \times \frac{W}{16}$	Patch Embedding	Patch Size	3×3, stride 2				
			Embed. Dim.	320			384	
		PoolFormer Block	Pooling Size	3×3, stride 1				
			MLP Ratio	4				
			# Block	6	12	18	18	24
4	$\frac{H}{32} \times \frac{W}{32}$	Patch Embedding	Patch Size	3×3, stride 2				
			Embed. Dim.	512			768	
		PoolFormer Block	Pooling Size	3×3, stride 1				
			MLP Ratio	4				
			# Block	2	4	6	6	8
Parameters (M)				11.9	21.4	30.8	56.1	73.4
MACs (G)				1.8	3.4	5.0	8.8	11.6

Experimental Results

ImageNet-1k Classification Results

General Arch.	Token Mixer	Outcome Model	Image Size	Params (M)	MACs (G)	Top-1 (%)
Convolutional Neural Netowrks	—	RSB-ResNet-18 [24, 59]	224	12	1.8	70.6
		RSB-ResNet-34 [24, 59]	224	22	3.7	75.5
		RSB-ResNet-50 [24, 59]	224	26	4.1	79.8
		RSB-ResNet-101 [24, 59]	224	45	7.9	81.3
		RSB-ResNet-152 [24, 59]	224	60	11.6	81.8
MetaFormer	Attention	ViT-B/16* [17]	224	86	17.6	79.7
		ViT-L/16* [17]	224	307	63.6	76.1
		DeiT-S [53]	224	22	4.6	79.8
		DeiT-B [53]	224	86	17.5	81.8
		PVT-Tiny [57]	224	13	1.9	75.1
		PVT-Small [57]	224	25	3.8	79.8
		PVT-Medium [57]	224	44	6.7	81.2
		PVT-Large [57]	224	61	9.8	81.7
	Spatial MLP	MLP-Mixer-B/16 [51]	224	59	12.7	76.4
		ResMLP-S12 [52]	224	15	3.0	76.6
		ResMLP-S24 [52]	224	30	6.0	79.4
		ResMLP-B24 [52]	224	116	23.0	81.0
		Swin-Mixer-T/D24 [36]	256	20	4.0	79.4
		Swin-Mixer-T/D6 [36]	256	23	4.0	79.7
		Swin-Mixer-B/D24 [36]	224	61	10.4	81.3
		gMLP-S [35]	224	20	4.5	79.6
		gMLP-B [35]	224	73	15.8	81.6
	Pooling	PoolFormer-S12	224	12	1.8	77.2
		PoolFormer-S24	224	21	3.4	80.3
		PoolFormer-S36	224	31	5.0	81.4
		PoolFormer-M36	224	56	8.8	82.1
		PoolFormer-M48	224	73	11.6	82.5

Table 2. **Performance of different types of models on ImageNet-1K classification.** All these models are only trained on the ImageNet-

Accuracy vs Model Size

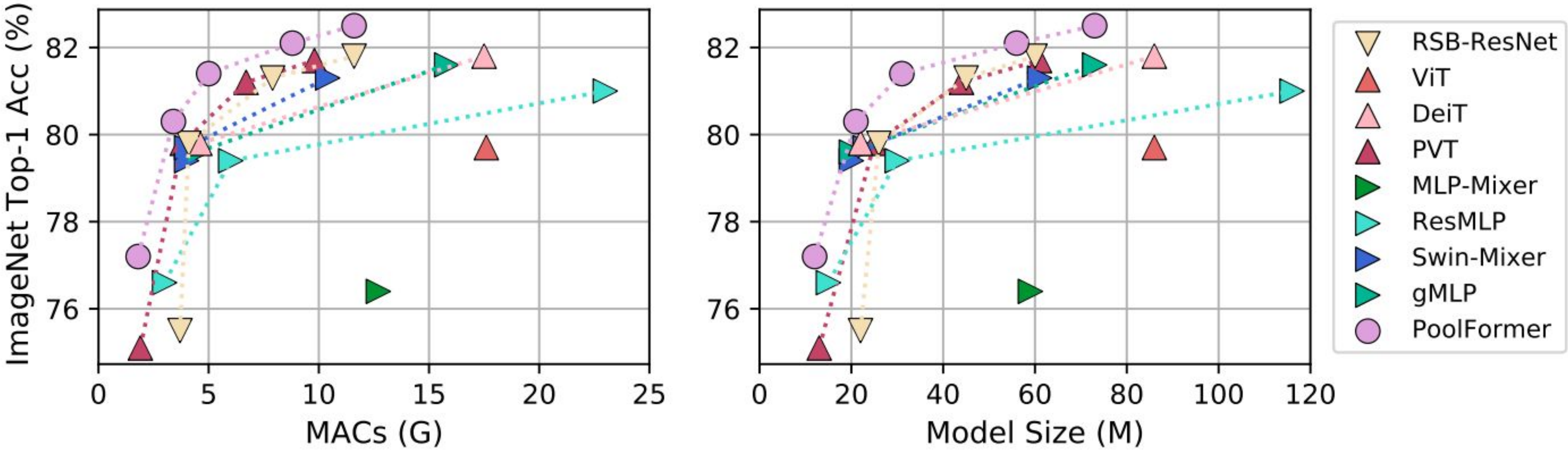


Figure 3. **ImageNet-1K validation accuracy *vs*. MACs/Model Size.** RSB-ResNet means the results are from "ResNet Strikes Back" [59] where ResNet [24] is trained with improved training procedure for 300 epochs.

COCO Object Detection Results

Backbone	RetinaNet 1×							Mask R-CNN 1×						
	Params (M)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params (M)	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet-18 [24]	21.3	31.8	49.6	33.6	16.3	34.3	43.2	31.2	34.0	54.0	36.7	31.2	51.0	32.7
PoolFormer-S12	21.7	36.2	56.2	38.2	20.8	39.1	48.0	31.6	37.3	59.0	40.1	34.6	55.8	36.9
ResNet-50 [24]	37.7	36.3	55.3	38.6	19.3	40.0	48.8	44.2	38.0	58.6	41.4	34.4	55.1	36.7
PoolFormer-S24	31.1	38.9	59.7	41.3	23.3	42.1	51.8	41.0	40.1	62.2	43.4	37.0	59.1	39.6
ResNet-101 [24]	56.7	38.5	57.8	41.2	21.4	42.6	51.1	63.2	40.4	61.1	44.2	36.4	57.7	38.8
PoolFormer-S36	40.6	39.5	60.5	41.8	22.5	42.9	52.4	50.5	41.0	63.1	44.8	37.7	60.1	40.0

Table 3. **Performance of object detection using RetinaNet, and object detection and instance segmentation using Mask R-CNN on COCO `val2017` [34].** 1× training schedule (*i.e.*12 epochs) is used for training detection models. AP^b and AP^m represent bounding box AP and mask AP, respectively.

ADE20K Semantic Segmentation Results

Backbone	Semantic FPN	
	Params (M)	mIoU (%)
ResNet-18 [24]	15.5	32.9
PVT-Tiny [57]	17.0	35.7
PoolFormer-S12	15.7	37.2
ResNet-50 [24]	28.5	36.7
PVT-Small [57]	28.2	39.8
PoolFormer-S24	23.2	40.3
ResNet-101 [24]	47.5	38.8
ResNeXt-101-32x4d [62]	47.1	39.7
PVT-Medium [57]	48.0	41.6
PoolFormer-S36	34.6	42.0
PVT-Large [57]	65.1	42.1
PoolFormer-M36	59.8	42.4
ResNeXt-101-64x4d [62]	86.4	40.2
PoolFormer-M48	77.1	42.7

Table 4. **Performance of Semantic segmentation on ADE20K [67] validation set.** All models are equipped with Semantic FPN [30].

Ablation Studies - Token Mixers

Ablation	Variant	Params (M)	MACs (G)	Top-1 (%)
Baseline	None (PoolFormer-S12)	11.9	1.8	77.2
Token mixers	Pooling $\rightarrow$ Identity mapping	11.9	1.8	74.3
	Pooling $\rightarrow$ Global random matrix* (extra 21M frozen parameters)	11.9	3.3	75.8
	Pooling $\rightarrow$ Depthwise Convolution [9, 38]	11.9	1.8	78.1
	Pooling size 3 $\rightarrow$ 5	11.9	1.8	77.2
	Pooling size 3 $\rightarrow$ 7	11.9	1.8	77.1
	Pooling size 3 $\rightarrow$ 9	11.9	1.8	76.8
Normalization	Modified Layer Normalization† $\rightarrow$ Layer Normalization [1]	11.9	1.8	76.5
	Modified Layer Normalization† $\rightarrow$ Batch Normalization [28]	11.9	1.8	76.4
	Modified Layer Normalization† $\rightarrow$ None	11.9	1.8	46.1
Activation	GELU [25] $\rightarrow$ ReLU [41]	11.9	1.8	76.4
	GELU $\rightarrow$ SiLU [18]	11.9	1.8	77.2
Other components	Residual connection [25] $\rightarrow$ None	11.9	1.8	0.1
	Channel MLP $\rightarrow$ None	2.5	0.2	5.7
Hybrid Stages	[Pool, Pool, Pool, Pool] $\rightarrow$ [Pool, Pool, Pool, Attention]	14.0	1.9	78.3
	[Pool, Pool, Pool, Pool] $\rightarrow$ [Pool, Pool, Attention, Attention]	16.5	2.5	81.0
	[Pool, Pool, Pool, Pool] $\rightarrow$ [Pool, Pool, Pool, SpatialFC]	11.9	1.8	77.5
	[Pool, Pool, Pool, Pool] $\rightarrow$ [Pool, Pool, SpatialFC, SpatialFC]	12.2	1.9	77.9

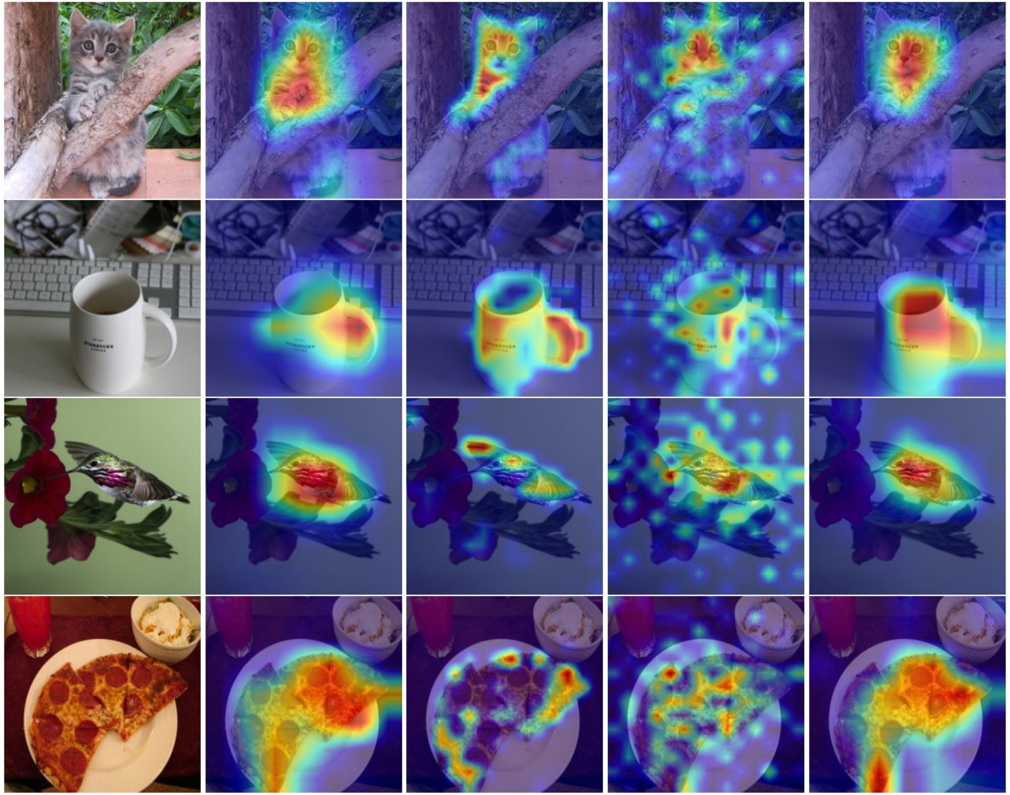
Table 5. **Ablation for PoolFormer on ImageNet-1K classification benchmark.** PoolFormer-S12 is utilized as the baseline to conduct

Ablation Studies - Hybrid Stages

Ablation	Variant	Params (M)	MACs (G)	Top-1 (%)
Baseline	None (PoolFormer-S12)	11.9	1.8	77.2
Token mixers	Pooling $\rightarrow$ Identity mapping	11.9	1.8	74.3
	Pooling $\rightarrow$ Global random matrix* (extra 21M frozen parameters)	11.9	3.3	75.8
	Pooling $\rightarrow$ Depthwise Convolution [9, 38]	11.9	1.8	78.1
	Pooling size 3 $\rightarrow$ 5	11.9	1.8	77.2
	Pooling size 3 $\rightarrow$ 7	11.9	1.8	77.1
	Pooling size 3 $\rightarrow$ 9	11.9	1.8	76.8
Normalization	Modified Layer Normalization† $\rightarrow$ Layer Normalization [1]	11.9	1.8	76.5
	Modified Layer Normalization† $\rightarrow$ Batch Normalization [28]	11.9	1.8	76.4
	Modified Layer Normalization† $\rightarrow$ None	11.9	1.8	46.1
Activation	GELU [25] $\rightarrow$ ReLU [41]	11.9	1.8	76.4
	GELU $\rightarrow$ SiLU [18]	11.9	1.8	77.2
Other components	Residual connection [25] $\rightarrow$ None	11.9	1.8	0.1
	Channel MLP $\rightarrow$ None	2.5	0.2	5.7
Hybrid Stages	[Pool, Pool, Pool, Pool] $\rightarrow$ [Pool, Pool, Pool, Attention]	14.0	1.9	78.3
	[Pool, Pool, Pool, Pool] $\rightarrow$ [Pool, Pool, Attention, Attention]	16.5	2.5	81.0
	[Pool, Pool, Pool, Pool] $\rightarrow$ [Pool, Pool, Pool, SpatialFC]	11.9	1.8	77.5
	[Pool, Pool, Pool, Pool] $\rightarrow$ [Pool, Pool, SpatialFC, SpatialFC]	12.2	1.9	77.9

Table 5. **Ablation for PoolFormer on ImageNet-1K classification benchmark.** PoolFormer-S12 is utilized as the baseline to conduct

Qualitative Results



Input RSB-ResNet-50 [59] DeiT-small [53] ResMLP-S24 [52] PoolFormer-S24

Conclusion

- Traditionally it is believed that attention is the key to success of the Transformers such as T2T-ViT, PVT, Swin, etc.
    - Attention is all you need

- In this work, the authors introduced a general Transformer architecture named MetaFormer by abstracting the attention layers
    - MetaFormer with simple pooling instead of attention delivers competitive performance on different vision tasks
    - MetaFormer is actually what you need for vision

Thanks

Any Questions?