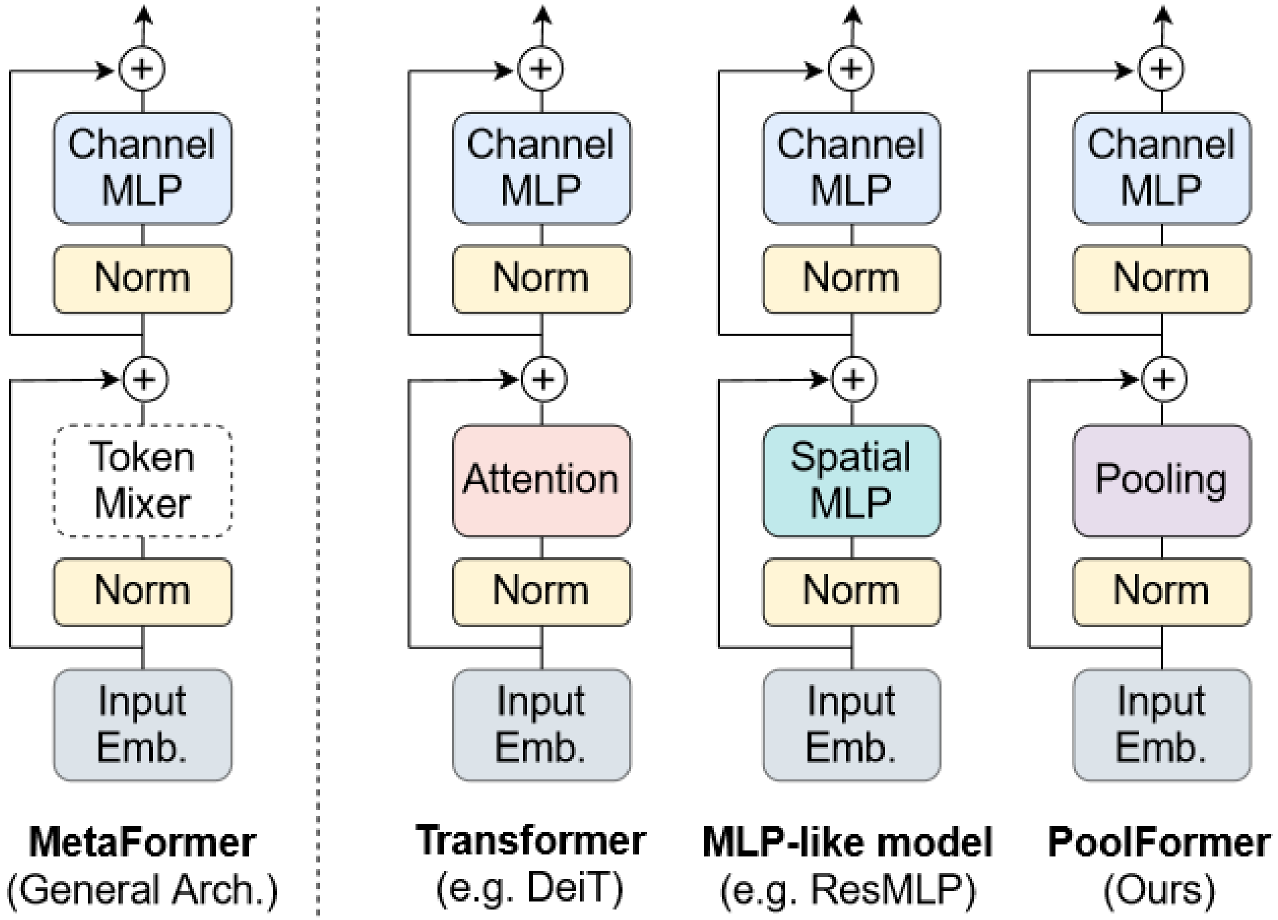
# MetaFormer is Actually What You Need for Vision

Weihao Yu Mi Luo Pan Zhou Chenyang Si Yichen Zhou,
Xinchao Wang Jiashi Feng Shuicheng Yan
Sea AI Lab National University of Singapore

# Outline

- analysis
  - ➢ attention-based token mixer module contributes most to their competence.
  - ➢ They can be replaced by spatial MLPs and the resulted models still perform quite well.
  - ➢ replace the attention module in transformers with an embarrassingly simple spatial pooling operator to conduct only the most basic token mixing.
- MetaFormer
- Experiments

# MetaFormer




embedding tokens

embedding dimension

$$X = \mathrm{InputEmb}(I), \quad X \in \mathbb{R}^{N \times C}$$

sub-block 1

$$Y = \mathrm{TokenMixer}(\mathrm{Norm}(X)) + X,$$

sub-block 2

$$Z = \sigma(\mathrm{Norm}(Y)W_1)W_2 + Y,$$

Activation Function

$$W_1 \in \mathbb{R}^{C \times rC}$$

$$W_2 \in \mathbb{R}^{rC \times C}$$

learnable parameters with MLP expansion ratio r

# PoolFormer

pool size

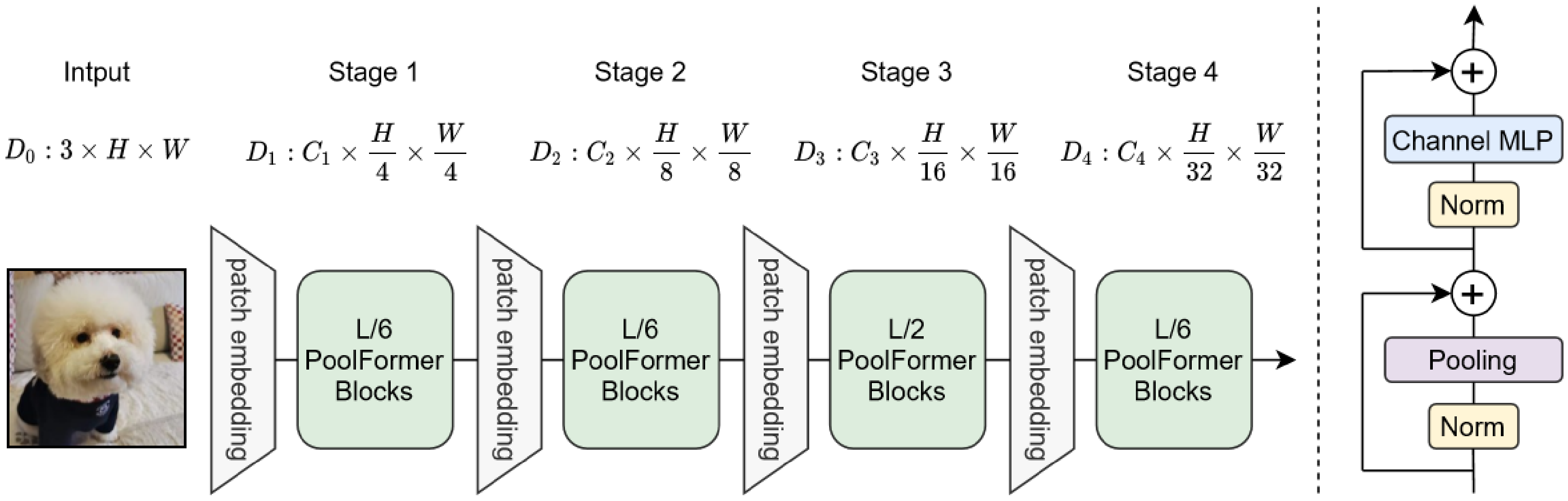
Pooling operator

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

**Algorithm 1** Pooling for PoolFormer, PyTorch-like Code

```
import torch.nn as nn

class Pooling(nn.Module):
    def __init__(self, pool_size=3):
        super().__init__()
        self.pool = nn.AvgPool2d(
            pool_size, stride=1,
            padding=pool_size//2,
            count_include_pad=False,
        )
    def forward(self, x):
        # [B, C, H, W] = x.shape
        return self.pool(x) - x
```




(a) Overal framework with L PoolFormer blocks

(b) PoolFormer block

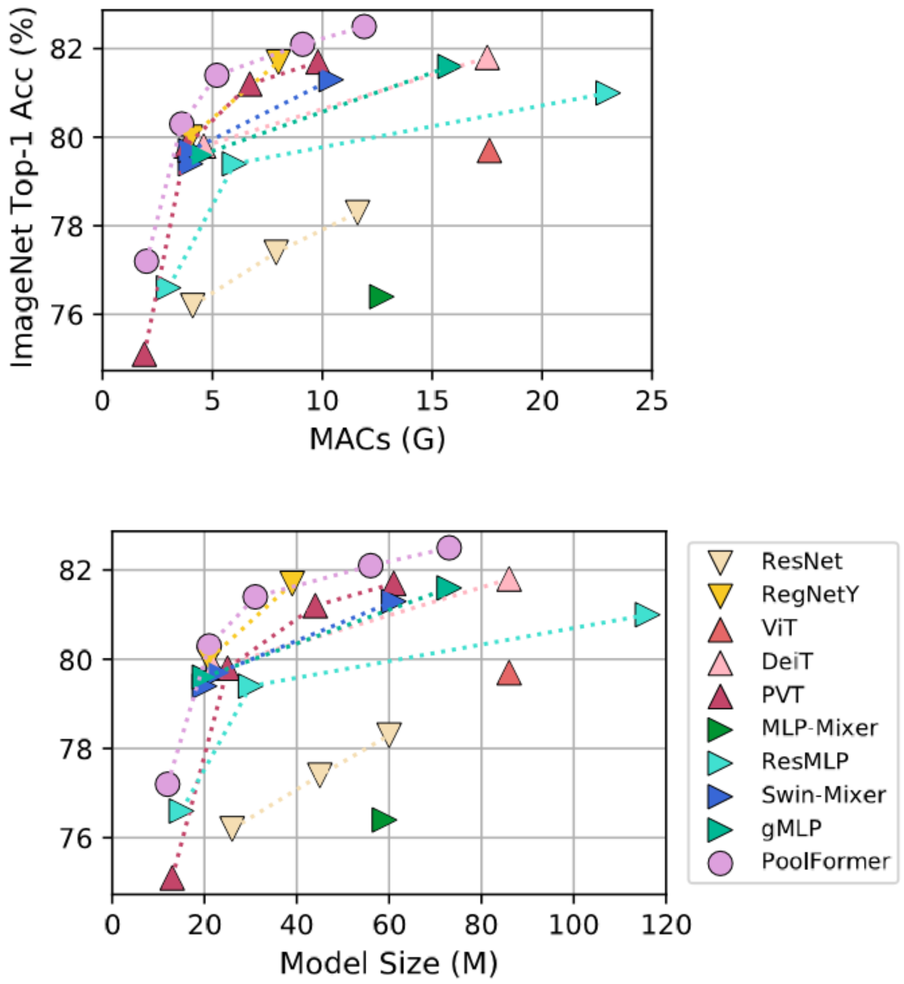
# PoolFormer

| Stage | # Tokens | Layer Specification | | PoolFormer S12 | PoolFormer S24 | PoolFormer S36 | PoolFormer M36 | PoolFormer M48 |
|---|---|---|---|---|---|---|---|---|
| 1 | $\frac{H}{4} \times \frac{W}{4}$ | Patch Embedding | Patch Size | $7 \times 7$, stride 4 | | | | |
| | | | Embed. Dim. | 64 | | | 96 | |
| | | PoolFormer Block | Pooling Size | $3 \times 3$, stride 1 | | | | |
| | | | MLP Ratio | 4 | | | | |
| | | | # Block | 2 | 4 | 6 | 6 | 8 |
| 2 | $\frac{H}{8} \times \frac{W}{8}$ | Patch Embedding | Patch Size | $3 \times 3$, stride 2 | | | | |
| | | | Embed. Dim. | 128 | | | 192 | |
| | | PoolFormer Block | Pooling Size | $3 \times 3$, stride 1 | | | | |
| | | | MLP Ratio | 4 | | | | |
| | | | # Block | 2 | 4 | 6 | 6 | 8 |
| 3 | $\frac{H}{16} \times \frac{W}{16}$ | Patch Embedding | Patch Size | $3 \times 3$, stride 2 | | | | |
| | | | Embed. Dim. | 320 | | | 384 | |
| | | PoolFormer Block | Pooling Size | $3 \times 3$, stride 1 | | | | |
| | | | MLP Ratio | 4 | | | | |
| | | | # Block | 6 | 12 | 18 | 18 | 24 |
| 4 | $\frac{H}{32} \times \frac{W}{32}$ | Patch Embedding | Patch Size | $3 \times 3$, stride 2 | | | | |
| | | | Embed. Dim. | 512 | | | 768 | |
| | | PoolFormer Block | Pooling Size | $3 \times 3$, 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) | | | | 2.0 | 3.6 | 5.2 | 9.1 | 11.9 |

# ImageNet Classification

**Dataset：** ImageNet-1k

| General Arch. | Token Mixer | Outcome Model | Image Size | Params (M) | MACs (G) | Top-1 (%) |
|---|---|---|---|---|---|---|
| Convolutional Neural Netowrks | — | ResNet-50 [22] | 224 | 26 | 4.1 | 76.2 |
| | | ResNet-101 [22] | 224 | 45 | 7.9 | 77.4 |
| | | ResNet-152 [22] | 224 | 60 | 11.6 | 78.3 |
| | | RegNetY-4GF [39] | 224 | 21 | 4.0 | 80.0 |
| | | RegNetY-8GF [39] | 224 | 39 | 8.0 | 81.7 |
| MetaFormer | Attention | ViT-B/16* [16] | 224 | 86 | 17.6 | 79.7 |
| | | ViT-L/16* [16] | 224 | 307 | 63.6 | 76.1 |
| | | DeiT-S [47] | 224 | 22 | 4.6 | 79.8 |
| | | DeiT-B [47] | 224 | 86 | 17.5 | 81.8 |
| | | PVT-Tiny [51] | 224 | 13 | 1.9 | 75.1 |
| | | PVT-Small [51] | 224 | 25 | 3.8 | 79.8 |
| | | PVT-Medium [51] | 224 | 44 | 6.7 | 81.2 |
| | | PVT-Large [51] | 224 | 61 | 9.8 | 81.7 |
| | Spatial MLP | MLP-Mixer-B/16 [45] | 224 | 59 | 12.7 | 76.4 |
| | | ResMLP-S12 [46] | 224 | 15 | 3.0 | 76.6 |
| | | ResMLP-S24 [46] | 224 | 30 | 6.0 | 79.4 |
| | | ResMLP-B24 [46] | 224 | 116 | 23.0 | 81.0 |
| | | Swin-Mixer-T/D24 [34] | 256 | 20 | 4.0 | 79.4 |
| | | Swin-Mixer-T/D6 [34] | 256 | 23 | 4.0 | 79.7 |
| | | Swin-Mixer-B/D24 [34] | 224 | 61 | 10.4 | 81.3 |
| | | gMLP-S [33] | 224 | 20 | 4.5 | 79.6 |
| | | gMLP-B [33] | 224 | 73 | 15.8 | 81.6 |
| | Pooling | PoolFormer-S12 | 224 | 12 | 2.0 | 77.2 |
| | | PoolFormer-S24 | 224 | 21 | 3.6 | 80.3 |
| | | PoolFormer-S36 | 224 | 31 | 5.2 | 81.4 |
| | | PoolFormer-M36 | 224 | 56 | 9.1 | 82.1 |
| | | PoolFormer-M48 | 224 | 73 | 11.9 | 82.5 |



# Object Detection and instance Segmentation

**Dataset： COCO**

Object Detection

| Model | Params (M) | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|
| ResNet-18 [22] | 21.3 | 31.8 | 49.6 | 33.6 | 16.3 | 34.3 | 43.2 |
| PoolFormer-S12 | 21.7 | 36.2 | 56.2 | 38.2 | 20.8 | 39.1 | 48.0 |
| ResNet-50 [22] | 37.7 | 36.3 | 55.3 | 38.6 | 19.3 | 40.0 | 48.8 |
| PoolFormer-S24 | 31.1 | 38.9 | 59.7 | 41.3 | 23.3 | 42.1 | 51.8 |
| ResNet-101 [22] | 56.7 | 38.5 | 57.8 | 41.2 | 21.4 | 42.6 | 51.1 |
| PoolFormer-S36 | 40.6 | 39.5 | 60.5 | 41.8 | 22.5 | 42.9 | 52.4 |

Instance Segmentation

| Model | Params (M) | $AP^b$ | $AP^b_{50}$ | $AP^b_{75}$ | $AP^m$ | $AP^m_{50}$ | $AP^m_{75}$ |
|---|---|---|---|---|---|---|---|
| ResNet-18 [22] | 31.2 | 34.0 | 54.0 | 36.7 | 31.2 | 51.0 | 32.7 |
| PoolFormer-S12 | 31.6 | 37.3 | 59.0 | 40.1 | 34.6 | 55.8 | 36.9 |
| ResNet-50 [22] | 44.2 | 38.0 | 58.6 | 41.4 | 34.4 | 55.1 | 36.7 |
| PoolFormer-S24 | 41.0 | 40.1 | 62.2 | 43.4 | 37.0 | 59.1 | 39.6 |
| ResNet-101 [22] | 63.2 | 40.4 | 61.1 | 44.2 | 36.4 | 57.7 | 38.8 |
| PoolFormer-S36 | 50.5 | 41.0 | 63.1 | 44.8 | 37.7 | 60.1 | 40.0 |

**Dataset： ADE20K**

Sementic Segmentation

| Model | Params (M) | mIoU (%) |
|---|---|---|
| ResNet-18 [22] | 15.5 | 32.9 |
| PVT-Tiny [51] | 17.0 | 35.7 |
| PoolFormer-S12 | 15.7 | 37.2 |
| ResNet-50 [22] | 28.5 | 36.7 |
| PVT-Small [51] | 28.2 | 39.8 |
| PoolFormer-S24 | 23.2 | 40.3 |
| ResNet-101 [22] | 47.5 | 38.8 |
| ResNeXt-101-32x4d [56] | 47.1 | 39.7 |
| PVT-Medium [51] | 48.0 | 41.6 |
| PoolFormer-S36 | 34.6 | 42.0 |
| PVT-Large [51] | 65.1 | 42.1 |
| PoolFormer-M36 | 59.8 | 42.4 |
| ResNeXt-101-64x4d [56] | 86.4 | 40.2 |
| PoolFormer-M48 | 77.1 | 42.7 |

# Ablation Studies

| Ablation | Variant | Params (M) | MACs (G) | Top-1 (%) |
| --- | --- | --- | --- | --- |
| Baseline | None (PoolFormer-S12) | 11.9 | 2.0 | 77.2 |
| Polling | Pooling → Identity mapping | 11.9 | 2.0 | 74.3 |
| | Pooling size 3 → 5 | 11.9 | 2.0 | 77.2 |
| | Pooling size 3 → 7 | 11.9 | 2.0 | 77.1 |
| | Pooling size 3 → 9 | 11.9 | 2.0 | 76.8 |
| Normalization | Group Normalization [55] → Layer Normalization [1] | 11.9 | 2.0 | 76.5 |
| | Group Normalization [55] → Batch Normalization [26] | 11.9 | 2.0 | 76.4 |
| Activation | GELU [23] → ReLU [38] | 11.9 | 2.0 | 76.4 |
| | GELU → SiLU [17] | 11.9 | 2.0 | 77.2 |
| Hybrid Stages | [Pool, Pool, Pool, Pool] → [Pool, Pool, Pool, Attention] | 14.0 | 2.1 | 78.3 |
| | [Pool, Pool, Pool, Pool] → [Pool, Pool, Attention, Attention] | 16.5 | 2.7 | 81.0 |
| | [Pool, Pool, Pool, Pool] → [Pool, Pool, Pool, SpatialFC] | 11.9 | 2.0 | 77.5 |
| | [Pool, Pool, Pool, Pool] → [Pool, Pool, SpatialFC, SpatialFC] | 12.2 | 2.1 | 77.9 |