Weekly Summary from 2022-8-16 to 2022-8-30

Abstract

1. 【2022-08-16 至 2022-08-18】和医疗器械方向校友见面，了解业界对医学人工智能问题的意向，国内这一方向校友分布，以及如何发现医学人工智能问题。
2. 【2022-08-16 至 2022-08-18】阅读 CVPR2022 论文 5 篇，主题为神经网络 backbone。论文列表如下：

Name	Paper	Code
A ConvNet for the 2020s	https://arxiv.org/abs/2201.03545	https://github.com/facebookresearch/ConvNeXt
Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs	https://arxiv.org/abs/2203.06717	https://github.com/megvii-research/RepLKNet https://github.com/DingXiaoH/RepLKNet-pytorch
MPViT : Multi-Path Vision Transformer for Dense Prediction	https://arxiv.org/abs/2112.11010	https://github.com/youngwanLEE/MPViT
Mobile-Former: Bridging MobileNet and Transformer	https://arxiv.org/abs/2108.05895	/
MetaFormer is Actually What You Need for Vision	https://arxiv.org/abs/2111.11418	https://github.com/sail-sg/poolformer

3. 【2022-08-16 至 2022-08-18】暑研答辩：基于流形学习的空间转录组细胞分割。
4. 【2022-08-16 至 2022-08-18】项目进展：基于脉冲神经网络的跨视角匹配。
5. 【2022-08-16 至 2022-08-18】项目进展：基于流形特征的粒子匹配。

Details

1. 【2022-08-16 至 2022-08-18】和医疗器械方向校友见面，了解业界对医学人工智能问题的意向，国内这一方向校友分布，以及如何发现医学人工智能问题。

谈话人：刘明宇（邦勤）；朱忠敬（千岸科技）

要点摘录：

1.1. 北京/长三角、深圳/珠三角两地医疗器械公司的差异：长三角地区更加欢迎新技术和新概念，珠三角地区侧重落地的利润；

1.2. 西交校友主要分布于长三角、珠三角等地，年龄范围广，有传统医疗器械领域，也有创业中，可帮忙推荐；

1.3. 了解企业业务，去医院不同科室了解，工业界和医院期望人工智能能替代医生不愿意做的工作，形成一个广阔的认知；本专业基础扎实，有不可替代性。

2. 【2022-08-16 至 2022-08-18】阅读 CVPR2022 论文 5 篇，主题为神经网络 backbone。

2.1. A ConvNet for the 2020s

收获：发掘“纯粹的”卷积网络的性能；使用最新的优化手段，改进 ResNeXt 的分组卷积和 ResNet 的 Bottle Neck 顺序，在 ImageNet 上达到 SOTA，参数量更少。

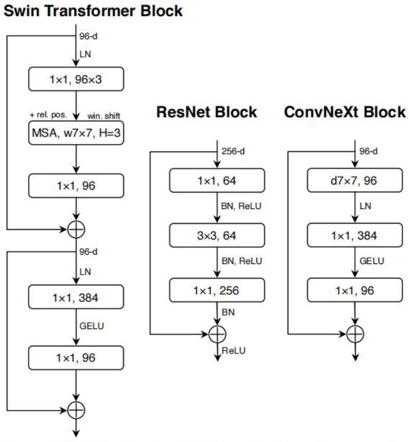


Figure 4. **Block designs** for a ResNet, a Swin Transformer, and a ConvNeXt. Swin Transformer's block is more sophisticated due to the presence of multiple specialized modules and two residual connections. For simplicity, we note the linear layers in Transformer MLP blocks also as “1×1 convs” since they are equivalent.

作者的意思，ConvNeXt 扩大感知域+合适优化手段是其进步的原因

2.2. Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs

收获：大的卷积核效果反而更好。

1. 超大卷积不但不涨点，而且还掉点？作者证明，超大卷积在过去没人用，不代表其现在不能用。人类对科学的认知总是螺旋上升的，在现代 CNN 设计（shortcut，重参数化等）的加持下，kernel size 越大越涨点！

2. 超大卷积效率很差？作者发现，超大 depth-wise 卷积并不会增加多少 FLOPs。如果再加点底层优化，速度会更快，31x31 的计算密度最高可达 3x3 的 70 倍！

3. 大卷积只能用在大 feature map 上？作者发现，在 7x7 的 feature map 上用 13x13 卷积都能涨点。

4. ImageNet 点数说明一切？作者发现，下游（目标检测、语义分割等）任务的性能可能跟 ImageNet 关系不大。

5. 超深 CNN（如 ResNet-152）堆叠大量 3x3，所以感受野很大？作者发现，深层小 kernel

模型有效感受野其实很小。反而少量超大卷积核的有效感受野非常大。

6. Transformers（ViT、Swin 等）在下游任务上性能强悍，是因为 self-attention（Query-Key-Value 的设计形式）本质更强？作者用超大卷积核验证，发现 kernel size 可能才是下游涨点的关键。

2.3. MPViT : Multi-Path Vision Transformer for Dense Prediction

收获：

密集的计算机视觉任务，如目标检测和分割，需要有效的多尺度特征表示，以检测或分类不同大小的物体或区域。Vision Transformer(ViT)构建了一个简单的多阶段结构（即精细到粗糙），用于使用单尺度 patch 的多尺度表示。然而 ViT 的变体专注于降低自注意的二次复杂度，较少关注构建有效的多尺度表示。

CoaT 通过使用一种 co-scale 机制，同时表示精细和粗糙的特征，允许并行地跨层注意，从而提高了检测性能。然而，co-scale 机制需要大量的计算和内存开销，因为它为基础模型增加了额外的跨层关注(例如，CoaT-Lite)。因此，对于 ViT 体系结构的多尺度特征表示仍有改进的空间。

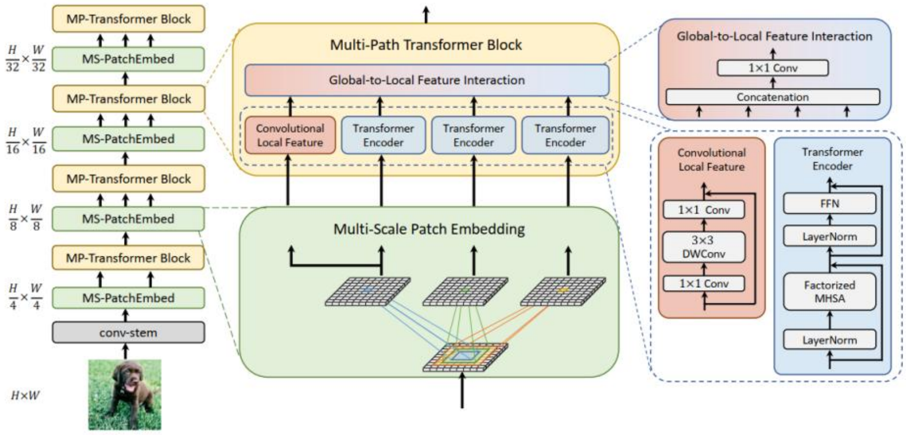

MP-Transformer Block
$\frac{H}{32}\times\frac{W}{32}$
MS-PatchEmbed
MP-Transformer Block
$\frac{H}{16}\times\frac{W}{16}$
MS-PatchEmbed
MP-Transformer Block
$\frac{H}{8}\times\frac{W}{8}$
MS-PatchEmbed
MP-Transformer Block
$\frac{H}{4}\times\frac{W}{4}$
MS-PatchEmbed
conv-stem
H×W
Multi-Path Transformer Block
Global-to-Local Feature Interaction
Convolutional Local Feature
Transformer Encoder
Transformer Encoder
Transformer Encoder
Multi-Scale Patch Embedding
Global-to-Local Feature Interaction
1×1 Conv
Concatenation
Convolutional Local Feature
1×1 Conv
3×3 DWConv
1×1 Conv
Transformer Encoder
FFN
LayerNorm
Factorized MHSA
LayerNorm

Figure 3. **Overview of Multi-Path Vision Transformer (MPViT).** MPViT consists of multi-scale patch embedding (MS-PatchEmbed) and multi-path transformer (MP-Transformer) blocks, which output features from each of the four stages for dense prediction tasks. Transformer encoders utilize factorized multi-head self-attention (MHSA) [65]. We omit the convoultional position encodings for simplicity.

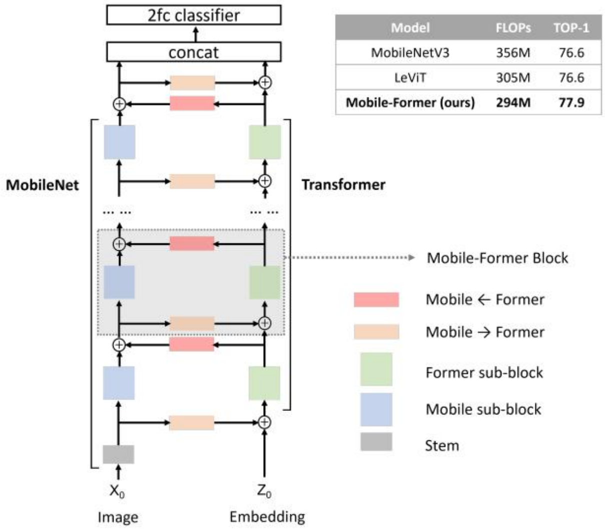
在多尺度层面上施加注意力

2.4. Mobile-Former: Bridging MobileNet and Transformer

收获：

在本文中，作者将设计范式从串联向并联转变，提出了一种新的 MobileNet 和 Transformer 并行化，并在两者之间建立双向桥接(见图)。将其命名为 Mobile-Former，其中 Mobile 指 MobileNet, Former 指 transformer。Mobile 以图像为输入堆叠 mobile block(或 inverted bottleneck)。它利用高效的 depthwise 和 pointwise 卷积来提取像素级的局部特征。前者以一些可学习的 token 作为输入，叠加 multi-head attention 和前馈网络(FFN)。这些 token 用于对图像的全局特征进行编码。

Mobile-Former 是 MobileNet 和 Transformer 的并行设计，中间有一个双向桥接。这种结构利用了 MobileNet 在局部处理和 Transformer 在全局交互方面的优势。并且该桥接可以实现局部和全局特征的双向融合。与最近在视觉 Transformer 上的工作不同，Mobile-Former 中的 Transformer 包含非常少的随机初始化的 token（例如少于 6 个 token），从而导致计算成本低。



Model	FLOPs	TOP-1
MobileNetV3	356M	76.6
LeViT	305M	76.6
Mobile-Former (ours)	**294M**	**77.9**

Figure 1. **Overview of Mobile-Former**, which parallelizes MobileNet [24] (on the left side) and Transformer [33] (on the right side). Different with vision transformer [8] that uses image patches to form tokens, the transformer in Mobile-Former takes *very few learnable tokens* as input that are randomly initialized. *Mobile* (refers to MobileNet) and *Former* (refers to transformer) are communicated by a bidirectional bridge, which is modeled by the proposed light-weight across attention. Best viewed in color.

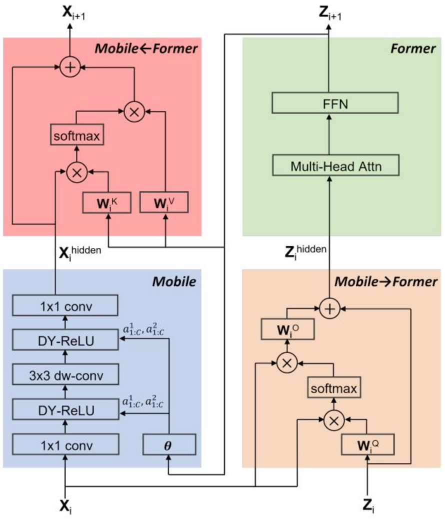


Figure 3. **Mobile-Former Block** that includes four modules: ***Mobile*** sub-block modifies inverted bottleneck block in [24] by replacing ReLU with dynamic ReLU [3]. ***Mobile→Former*** uses light-weight cross attention to fuse local features into global features. ***Former*** sub-block is a standard transformer block including multi-head attention and FFN. Note that the output of *Former* is used to generate parameters for dynamic ReLU in *Mobile* sub-block. ***Mobile←Former*** bridges from global to local features.

同时从嵌入和图像中学习，并且学习过程是相互交叉的

2.5. MetaFormer is Actually What You Need for Vision

收获：研究了ViT结构和类MLP结构的模型，将两者中相同的部分提取出来，组成了MetaFormer结构，并指出两者的性能都得左侧为MetaFormer结构，MetaFormer中的Token Mixer模块用于混合多个token之间的信息。该模块在类ViT结构的模型中对应为Attention模块（例如DeiT），而在类MLP结构的模型中对应SpatialMLP模块（例如GMLP、ResMLP）。

作者指出 MetaFormer 结构为类 ViT 和类 MLP 模型性能的主要来源，为了验证这一点，作者将 MetaFormer 结构中的 Token Moixer 模块替换为恒等映射（其实就是一个大卷积模型，一个单通道的大卷积处理一个特征图（token）），模型在 ImageNet 上的 top-1 准确率仍可达 74.3%。此外，作者通过实验发现去除 Norm、Channel MLP、Shortcut 中的任意一个模块，模型均难以收敛，由此验证了 MetaFormer 中的结构缺一不可。

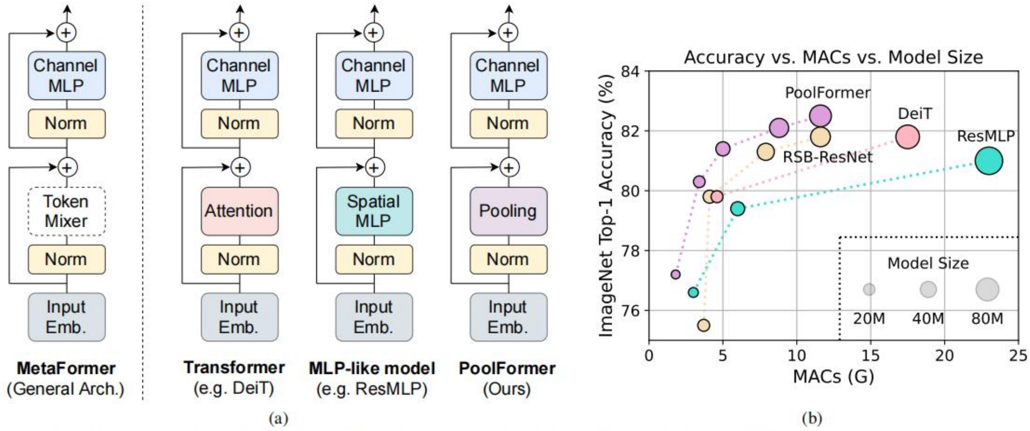
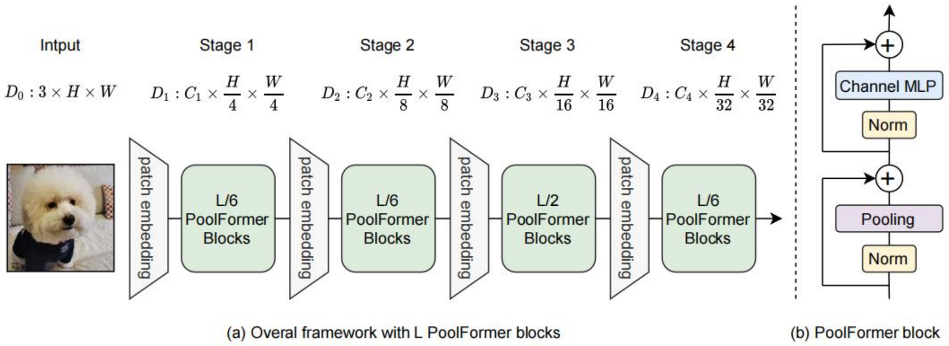


Figure 1. **MetaFormer and performance of MetaFormer-based models on ImageNet-1K validation set.** As shown in (a), we present *MetaFormer* as a general architecture abstracted from Transformers [56] by not specifying the token mixer. When using attention/spatial MLP as the token mixer, MetaFormer is instantiated as Transformer/MLP-like models. We argue that the competence of Transformer/MLP-like models primarily stems from the general architecture MetaFormer instead of the equipped specific token mixers. To demonstrate this, we exploit an embarrassingly simple non-parametric operator, *pooling*, to conduct extremely basic token mixing. Surprisingly, the resulted model *PoolFormer* consistently outperforms the well-tuned vision Transformer [17] baseline (DeiT [53]) and MLP-like [51] baseline (ResMLP [52]) as shown in (b), which well supports that MetaFormer is actually what we need to achieve competitive performance. RSB-ResNet in (b) means the results are from "ResNet Strikes Back" [59] where ResNet [24] are trained with improved training procedure for 300 epochs.


Intput
$D_0 : 3 \times H \times W$
Stage 1
$D_1 : C_1 \times \frac{H}{4} \times \frac{W}{4}$
Stage 2
$D_2 : C_2 \times \frac{H}{8} \times \frac{W}{8}$
Stage 3
$D_3 : C_3 \times \frac{H}{16} \times \frac{W}{16}$
Stage 4
$D_4 : C_4 \times \frac{H}{32} \times \frac{W}{32}$
patch embedding
L/6 PoolFormer Blocks
patch embedding
L/6 PoolFormer Blocks
patch embedding
L/2 PoolFormer Blocks
patch embedding
L/6 PoolFormer Blocks
Channel MLP
Norm
Pooling
Norm

(a) Overal framework with L PoolFormer blocks

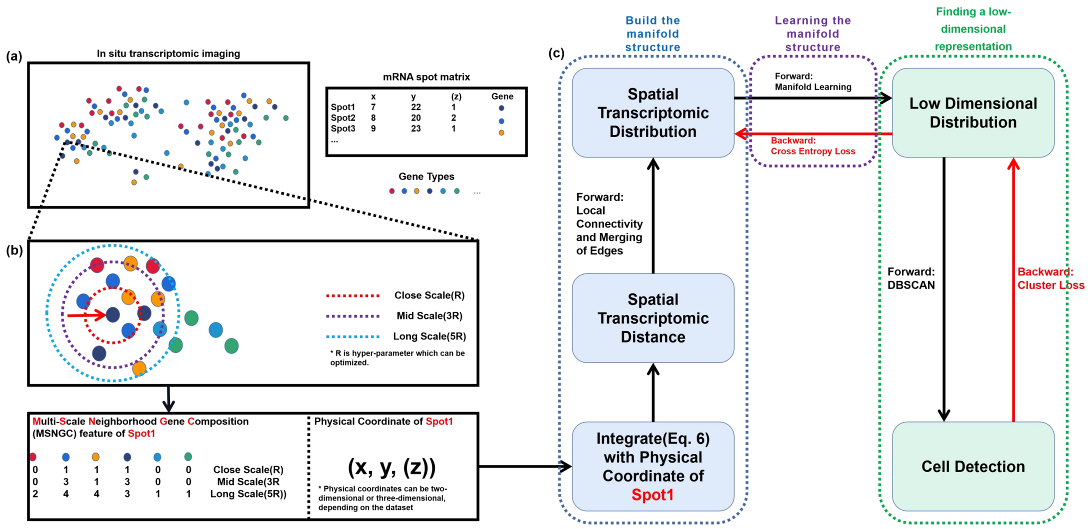
(b) PoolFormer block

Figure 2. **(a) The overall framework of PoolFormer.** Similar to [24, 36, 57], PoolFormer adopts hierarchical architecture with 4 stages. For a model with L PoolFormer blocks, stage [1, 2, 3, 4] have [L/6, L/6, L/2, L/6] blocks, respectively. The feature dimension D_i of stage i is shown in the figure. **(b) The architecture of PoolFormer block.** Compared with Transformer block, it replaces attention with extremely simple non-parametric operator, pooling, to conduct only basic token mixing.

作者在 MetaFormer 的基础上引入了 PoolFormer 结构，PoolFormer 通过平均池化融合多个 Token 之间的信息，相比于 Attention 和 SpatialMLP，池化操作不会引入额外的参数，并且计算量更小

3. 【2022-08-16 至 2022-08-18】暑研答辩：基于流形学习的空间转录组细胞分割。

思路：融合多尺度邻域基因信息和空间信息构建空间转录组局部连接的流形结构，通过数据降维在欧几里得空间投影，并进行密度聚类，在优化的时候同时考虑局部连接和每一个聚类簇的连接。


(a)
In situ transcriptomic imaging
mRNA spot matrix
x y (z) Gene
Spot1 7 22 1
Spot2 8 20 2
Spot3 9 23 1
...
Gene Types
...
(b)
Close Scale(R)
Mid Scale(3R)
Long Scale(5R)
* R is hyper-parameter which can be optimized.
Multi-Scale Neighborhood Gene Composition (MSNGC) feature of Spot1
0 1 1 1 0 0 Close Scale(R)
0 3 1 3 0 0 Mid Scale(3R
2 4 4 3 1 1 Long Scale(5R))
Physical Coordinate of Spot1
(x, y, (z))
* Physical coordinates can be two-dimensional or three-dimensional, depending on the dataset
(c)
Build the manifold structure
Learning the manifold structure
Finding a low-dimensional representation
Spatial Transcriptomic Distribution
Forward: Manifold Learning
Backward: Cross Entropy Loss
Low Dimensional Distribution
Forward: Local Connectivity and Merging of Edges
Spatial Transcriptomic Distance
Forward: DBSCAN
Backward: Cluster Loss
Integrate(Eq. 6) with Physical Coordinate of Spot1
Cell Detection

ST-UMAP spatial transcriptome cell segmentation algorithm.(a) Space transcriptomic data structure. Each spatial transcriptome sampling site has both physical coordinates and gene tags, and the physical coordinates can be two-dimensional or three-dimensional.(b) Definition of the spatial transcriptome distance. MSNGC feature is a multi-scale gene position feature that counts the number of adjacent genes at different scales, and represents the gene relationship between the sampling point and the surrounding sampling point. The physical coordinate is the spatial location information of each sampling point. Spatial transcriptomic distances were constructed by merging MSNGC features with physical coordinates.(c) End-to-end based on manifold learning spatial transcriptome cell segmentation framework. First, the spatial transcriptomic manifolds were constructed. A fully connected manifold is constructed by considering the local connectivity and merging edges. The distribution of the cluster-spatially sampled clusters, defined by the Euclidean distance, approximates the spatial transcriptome manifolds by means of manifold learning. Cell segmentation results were obtained using density clustering in the clustering space. Fusion cross-entropy and clustering analysis metrics are used as error propagation for trained manifold approximators.

4. 【2022-08-16 至 2022-08-18】项目进展：基于脉冲神经网络的跨视角匹配。

收获：将训练梯度解耦为全局梯度和局部梯度，使用元学习的方式学习神经元的激活参数，能够端到端训练一个混合神经网络，用于跨视角匹配问题。

5. 【2022-08-16 至 2022-08-18】项目进展：基于流形特征的粒子匹配。

收获：对于来自于温和流场的粒子匹配，计算每一个粒子的局部连接流形，基于同帧的“邻域”粒子分布构造粒子级的特征；因为采样的质量（数目和概率分布）会影响置信度，我希望通过贝叶斯形式将迭代过程展开为一个一阶的 MC chain，即显变量是粒子流形的概率分布，隐变量是置信度概率，来考虑跨帧信息或绕开粒子突现的问题。