**“A ConvNet for the 2020s”论文阅读报告**

**姓名：陈鸿绪 学号：PB21000224 日期: 6.29.2024**

1. **背景简介：**

Transformer[1]架构的提出对自然语言处理（NLP）领域带来了深远的影响，Transformer架构摒弃了传统的循环神经网络（RNN）或长短时记忆网络（LSTM）结构，引入了全新的自注意力机制（Self-Attention Mechanism），这一创新使得文本数据的处理更为高效，Transformer在处理长序列时能够更好地捕捉长距离依赖关系，从而克服了传统RNN在处理长序列时存在的梯度消失和梯度爆炸问题。

Transformer架构在NLP领域的革命性突破让计算机视觉（CV）领域也开始了新架构的尝试，以基于自注意力机制架构的Vision Transformer[2]（ViT）成功地将Transformer引入到CV，而在这之前，卷积神经网络（CNN）在计算机视觉任务中一直占据主导地位。然而ViT的出现证明了纯Transformer架构在图像分类等任务上也能表现出色，直接挑战了CNN的统治地位。在研究人员ViT的研究基础上陆续提出了Transformer in Transformer（TNT）、Data-efficient image Transformers[3]（DeiT）、Swing Transformer[4]（SwinT，注意后续其与Swin-T的区别）等多种改进的ViT变体。其中拥有可变感受野、移位窗口自注意力机制、线性计算复杂度的SwinT以其巧妙的设计和在ImageNet数据集上的Sota（提出的时候）水平再次证明了自注意力机制的强大。在Transformer的冲击下的2020s，卷积神经网络似乎尽显颓势。

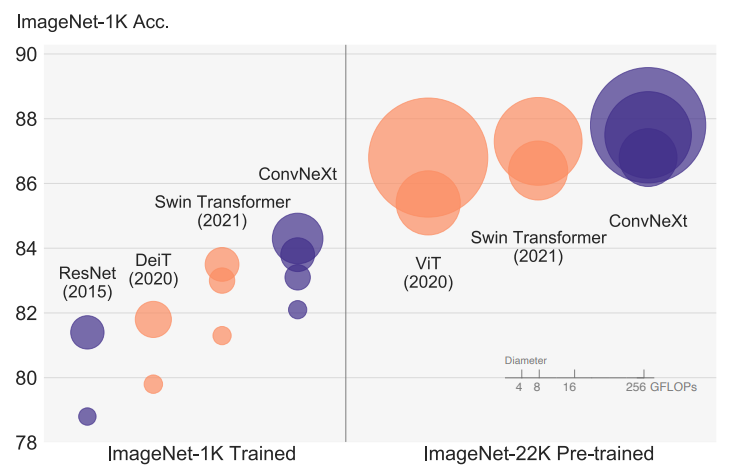
1. **问题提出：**

在计算机视觉领域，卷积的统治地位并非偶然。滑动窗口策略本质上与视觉处理的需求相契合，归纳偏置和平移不变性特质使得CNN在该领域的应用中极为适宜。尽管ViT曾面临挑战，即普通ViT的感受野是固定的，不能关注到一个patch中的内部信息，但这些困难通过引入类似卷积中滑动窗口策略的Swin Transformer得到了解决。这一事实进一步表现出卷积在视觉处理中的重要性，同时也启发我们思考如何将Transformer的优势与卷积网络相结合。基于Transformer的尝试往往伴随复杂的系统设计和非常大的算力代价，是否可以将这些设计思想移植到结构简单的CNN网络上，因为我们知道CNN之所以会显现颓势主要是因为Transformer基于其多头自注意力机制的优越的尺度行为。

在经过以上分析后，我们自然会考虑架构简单的CNN真的是弱于Transformer，是否可以得到比SwinT性能更加优越纯卷积神经网络。

1. **论文简介：**

Facebook AI Research和UC Berkeley的联合论文“A ConvNet for the 2020s”[5]探究了卷积神经网络的极限，这篇文章发表在了2022年CVPR会议上。该工作提出了ConvNeXt纯卷积网络，其在设计网络时想法理念对标于Swin Transformer，作者在此实验中发现了几个导致性能差异的关键组件。实验证明，ConvNeXt相比SwinT拥有更快推理速度与更高准确率，在ImageNet 22K上ConvNeXt-XL达到了87.8%的准确率（当时的Sota），从而证明了纯CNN结构并没有达到它的瓶颈。下面我将具体阐述论文的提出方法、实验结果与分析。

**图1. ImageNet-1K classification results**

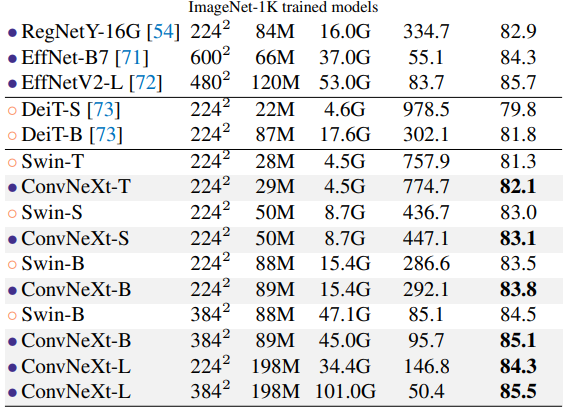
1. **提出方法:**

作者使用ViT的训练策略重新训练得到ResNet50[6]模型，其效果显著优于原始结果，所以将该结果认定为后续实验的baseline。ConvNeXt设计的主要创新点分为下面五个：

1. **Macro Design:** 对于SwinT设计遵循了过去ConvNets的设计规则，即使用了multi-stage design，所以为了对标SwinT设计，ConvNeXt的设计在这部分着重关注两个部分：stage compute ratio和 “stem cell” structure。
2. Stage compute ratio：原始ResNet50的block堆叠次数为(3, 4, 6, 3)，比例近似为1:1:2:1。参考Swin-T的堆叠block次数比例为1:1:3:1，Swin-L的比例为1:1:9:1，明显在SwinT在设计时其stage3的block占比相对其它stage较大，所以作者据此将原始ResNet50的堆叠次数改变成为(3, 3, 9, 3)，FLOPs与Swin-T相似。最后模型准确率从78.8%提升至79.4%。
3. Changing stem to “Patchify”: 在之前的卷积神经网络中，一般最初下采样stem一般使用一个7×7卷积核、stride为2的卷积层加上strade为2的Maxpooling下采样组成，该stem可以使得完成4×的下采样。但是ViT采取了更加激进的patchify方案，这对应着更大的卷积核，且相邻窗口没有交集（SwinT也采取了该方案）。对标SwinT，作者将stem替换成了patchify方案，替换之后准确率从79.4%上升到79.5%，且FLOPs也略微降低。
4. **ResNeXt-ify：**这部分作者主要运用了ResNeXt[7]的思想，即采用组卷积（卷积核被分成了不同组），采用组卷积后ResNeXt做到了比原始ResNet更好的FLOPs与accuracy之间的平衡。作者采用了深度可分离卷积设计方案（即组卷积在group数和channel数相同的特殊情况），作者采取这种动机在于他们认为组卷积某种程度上类似于自注意力机制的加权和。这种设计有效减少了网络的FLOPs。同时网络宽度也改变成了与Swin-T相同的通道数(64到96)，最后这些改变使得模型准确率上升到80.5%。
5. **Inverted Bottleneck：**Inverted Bottleneck指的是MLP block的隐藏层维度比输入向量维度宽4倍，是Transformer采用重要的一个设计。作者在模型设计中引入了Inverted Bottleneck，最后在小模型上准确率从80.5%上升至80.6%，但在大模型（如ResNet-200 / Swin-B regime），这步使得准确率得到较为显著的提升，从81.9%提升至82.6%同时也减少了FLOPs。
6. **Large Kernel Sizes：**过去ConvNets中曾使用过较大的卷积核尺寸，但VGGNet[8]普及了堆叠3×3小尺寸卷积层，所以现代GPU对3×3卷积做了很多的优化。考察SwinT的自注意力块中引入的局部窗口，其大小为7×7，这远大于3×3卷积核大小，所以我们需要重新审视ConvNets中使用大卷积层的做法。
7. 上移**depthwise conv layer：**为了探索大尺寸卷积核，作者将模型向上移动**depthwise**卷积层，这是在Transformers中较为明显类似的设计，因为MSA block也被放置在MLP层之前。这样改动准确率下降到79.9%，其FLOPs略微减少。
8. 加大卷积尺寸：作者接着将**depthwise卷积层的卷积核增大至7×7，作者在尝试其它尺寸时发现7达到了饱和性能。准确率从原本3×3卷积核的79.9%上升到7×7卷积核的80.6%。**
9. **Micro Design：**这部分主要关注微观尺寸上模型的变化，即这些探索主要聚焦于网络层，着重关注激活函数以及normalization层的选择。
10. 替换激活函数ReLU为GELU：在transformer的原始论文中，ReLU作为激活函数。但在BERT、OpenAI的GPT-2、ViTs模型以及先进的transformer模型中，高斯误差线性单元（GULE，可以看作ReLU的变体）得以应用。所以作者采取GELU代替ReLU，尽管准确率保持一致。
11. 更少的激活函数：考虑到Transformer中并不是每一个模型都接有激活函数，所以作者采取适当减少激活函数的方案，发现准确率从80.6%上升至81.3%。
12. 更少的Normalization层：考虑到Transformer中，Normalization层的使用有限，所以作者也减少了ConNeXT Block中的Normalization层的使用，只对depthwise卷积层后的Normalization层保留，最终准确率上升到了81.4%。超过了Swin-T。
13. 替代Batch Normalization（BN）为Layer Normalization（LN）：考虑到Transformer中基本使用的都为LN，作者将BN替换成了LN，发现准确率小幅度提升到了81.5%。
14. 分离出下采样层：考虑SwinT设计采用了Patch Merging操作，作者为ConvNext单独设计了下采样层，即LN后接2×2卷积层，进行了此设计后准确率提升至82.0%。
15. **实验结果及分析：**

实验训练的模型根据数据集分为了两类，一类是直接在ImageNet-1K上进行训练，另一类是通过在ImageNet-22K上预训练后再通过在ImageNet-1K上微调的模型。对上面两类模型在ImageNet-1K上进行分类任务准确率预测。同时也进行了ConvNeXt vs ViT的消融实验。最后将ConvNeXt运用在特定下游任务上评估其性能。

1. **ImageNet-1K（在ImageNet-1K上训练的若干模型）**



**表1：ImageNet-1K trained models准确率**

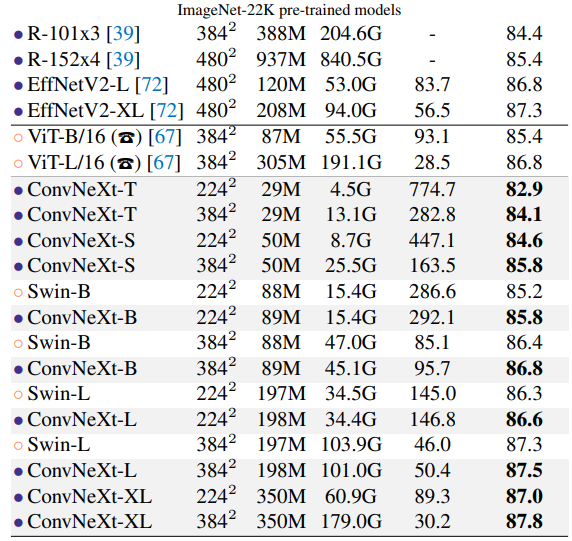
作者在上面表1展示了当时最新的Transformer变体（DeiT和SwinT）以及来自架构搜索的两种ConvNet（RegNets、EfficientNets[9]和EfficientNetsV2[10]）的结果对比，注意上述参数量相近的之间进行比较。ConvNeXt在准确性与计算量权衡方面，以及推理吞吐量方面，与两个强大的ConvNet基线（RegNet[11]和EfficientNet）相比具有一定竞争力。

此外，ConvNeXt全面超越具有相似复杂度的Swin Transformer，甚至有时其优势非常显著（例如，ConvNeXt-T比Swin-T高出0.8%的准确率）。然而ConvNeXt没有使用诸如移位窗口或相对位置偏差等专用模块，同时相比之下，ConvNeXt吞吐量也有所提高。

特别地，我们需要注意结果中ConvNeXt-B在测试图像在384×384分辨率下，它比Swin-B高出0.6%（84.5%到85.1%），但推理吞吐量高出12.5%（85.1到95.7图像/秒）。当分辨率从224×224增加到384×384时，ConvNeXt-B相对于Swin-B的浮点运算/吞吐量优势会变得更大。此外，当我们进一步扩展到ConvNeXt-L时，可以从上表观察到结果提高到85.5%。

1. **ImageNet-22K（在ImageNet-22K上预训练的若干模型）**

****



**表2：ImageNet-22K pre-trained models准确率**

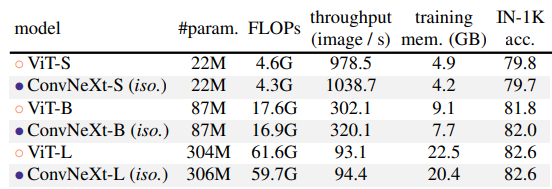
上表展示了使用ImageNet-22K预训练微调后的模型结果。学界普遍认为vision Transformers的归纳偏差较少，因此在更大规模的数据集上进行预训练时，性能会比基于纯卷积的ConvNets更好。

然而事实上，上表结果表明，当使用大型数据集进行预训练时，作者所设计的卷积神经网络并不逊色于基于Transformer的模型，ConvNeXts的性能仍然与参数量级一致的SwinT相当或更好，并且吞吐量略高。

此外，我们注意ConvNeXt-XL模型在图像384×384的分辨率下达到了87.8%的准确率，相比ConvNeXt-L的87.5%有了显著的改进，这证明了ConvNeXts是可扩展的架构。 在ImageNet-1K上，EfficientNetV2-L通过一系列模块配置达到了顶级性能。然而，通过模型在ImageNet-22K上的预训练，ConvNeXt能够超越EfficientNetV2（87.8%超出87.3%），这进一步证明了大规模训练对ConvNeXts性能表现的重要性。

1. **ConvNeXts与ViT消融对比实验**

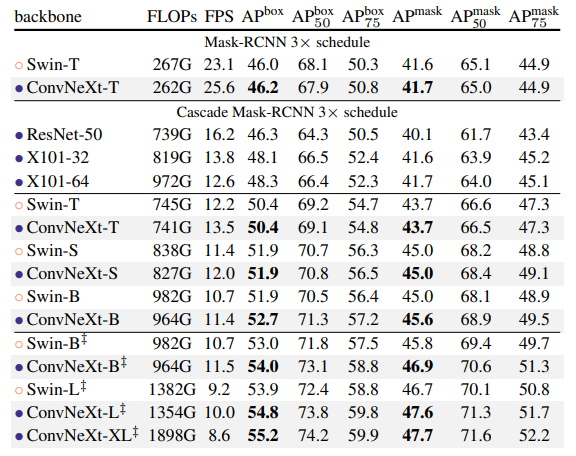
作者希望在这个部分探究ConvNeXt块设计是否可泛化到ViT风格的等轴架构中。所以作者构建了ConvNeXt-S/B/L，使用与ViT-S/B/L（384/768/1024）相同的特征维度，最后得到了在ImageNet-1K数据集的若干结果，如下表所示：



**表3. Comparing isotropic ConvNeXt and ViT**

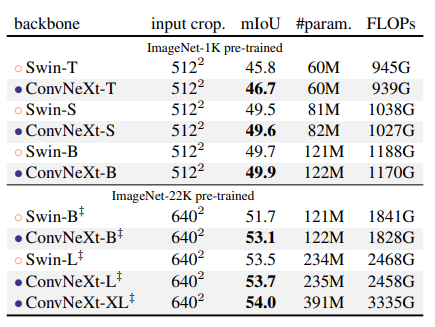
由上表，我们可以看见相同维度的ConvNeXt和ViT相比具有较强的竞争力，ConvNeXt-L和ConvNeXt-B在IN-1K acc上都不比相应的ViT-L与ViT-B低，且相同维度ConvNeXt训练占用显存是比ViT小的。所以可以认为ConvNeXt通常可以与ViT相媲美，这表明当用于非层次化模型时，ConvNeXt块设计具有竞争力，ConvNeXt块设计是可以泛化到ViT风格的等轴架构中的。

1. **下游任务表现**



**表4. COCO object detection and segmentation results**

可以从上表看出对于COCO物体检测和分割任务，相同维度而言，基于ConvNeXt的模型测试指标几乎都是相当或者优于基于SwinT模型的测试指标的。当扩大到更大的模型（在ImageNet-22K上预训练的ConvNeXt-B/L/XL）时，可以看出ConvNeXt在很多情况下是显著优于Swin Transformer（例如，+1.0 AP）。



**表5. ADE20K validation results**

作者还使用UperNet[12]在ADE20K语义分割任务上评估了ConvNeXt主干网络。所有模型变体都进行了160K次迭代训练，批次大小为16。在表5中，可以看出多尺度测试下的验证mIoU指标。上表显示ConvNeXt模型可以在不同模型容量下实现具有竞争力的性能，进一步验证了ConvNeXt模型架构设计的有效性。

1. **与课内所学的关联&阅读收获**

在深度学习导论课程中，我们学习了卷积神经网络和基于Transformer架构的网络模型，也知道了在学界开始逐渐开始形成一种刻板印象：认为传统卷积已经在CV领域输给基于自注意力机制的Transformer。这篇论文工作的出现说明了纯卷积网络其实还没有达到性能的尽头，只要它设计得当，它仍然和Transformer架构有相当的竞争力。

通过学习这篇论文的工作，我认识到在这个Transformer架构盛行的时代，需要重新审视传统卷积，纯卷积神经网络不仅拥有其简单直观的架构，还拥有仍然没有被完全开发的强大性能。同时通过论文中在大规模数据集（如ImageNet-22K）上进行预训练的实验，我明白了大规模预训练对于提高模型性能的重要性。除了架构设计外，论文还强调了训练策略对于提高模型性能的重要性。通过采用更长的预热周期和更精细的训练设置，ConvNeXt能够在相同的计算资源下获得更好的性能。

**参考文献**

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[3] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. PMLR, 2021: 10347-10357.

[4] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[5] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[7] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[9] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. PMLR, 2019: 6105-6114.

[10] Tan M, Le Q. Efficientnetv2: Smaller models and faster training[C]//International conference on machine learning. PMLR, 2021: 10096-10106.

[11] Radosavovic I, Kosaraju R P, Girshick R, et al. Designing network design spaces[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10428-10436.

[12] Xiao T, Liu Y, Zhou B, et al. Unified perceptual parsing for scene understanding[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 418-434.