# ConvNeXt 网络的设计与实验

**摘要**

本文针对近年来卷积神经网络与视觉 Transformer 的研究趋势，重点探讨了 ConvNeXt 网络的设计理念与性能表现。ConvNeXt 通过对 ResNet 结构的现代化改进，实现了接近甚至超过 Transformer 性能的表现，同时保留了卷积网络的高效性与推理速度。在在本文中，我们在 CIFAR-100 数据集上复现并探讨了 ConvNeXt 网络的特性，并结合实验结果与 ViT 和 ResNet-1001 进行了对比分析。

**关键词：** ConvNeXt；卷积神经网络；深度学习；特征提取；图像分类

## 1 引言

随着深度学习的发展，卷积神经网络（CNN）在图像分类、目标检测和语义分割等任务中取得了显著进展。然而，近年来基于视觉 Transformer（ViT）的模型由于其全球特征提取能力而引发了广泛关注，逐渐成为计算机视觉领域的研究热点。

尽管 ViT 展现出了强大的性能，其高昂的计算代价和对大规模训练数据的依赖使其难以在部分场景中推广。为此，论文《ConvNeXt》中提出了一种现代化改造的卷积神经网络架构，通过引入与 ViT 类似的设计元素（如深度可分离卷积、LayerNorm 等），在保留卷积高效性的同时显著提升了模型性能。

本文旨在对 ConvNeXt 的设计与实现进行深入分析。我们从其网络架构设计、复现细节与实验结果三个层面展开研究，并在 CIFAR-100 数据集上验证其性能，分析其与 ViT 和 ResNet-101 的对比结果。本文的主要贡献如下：

1. 系统性地分析了 ConvNeXt 的模块设计及其改进点；

2. 针对 CIFAR-100 数据集，提出了一种轻量化的 ConvNeXt 变体并优化了训练流程；

3. 对比分析了 ConvNeXt、ViT 和 ResNet 在分类任务上的性能差异，为卷积网络的现代化发展提供了新的视角。

## 2 相关工作

### 2.1 传统卷积神经网络

传统卷积神经网络通过堆叠卷积层与池化层来提取图像的局部特征。经典模型如 LeNet、AlexNet、VGG 和 ResNet 等为计算机视觉领域奠定了重要基础。其中，ResNet 通过引入残差连接解决了深层网络的梯度消失问题，显著提高了网络的训练效率和性能 [2]。

然而，这些模型通常缺乏对全局上下文信息的提取能力，无法在某些复杂视觉任务中达到最优表现。为了应对这一问题，研究者们开始探索 ViT 等全局特征提取方法。

### 2.2 视觉 Transformer

ViT 通过自注意力机制对全局特征进行建模，展现了优于 CNN 的性能。然而，其计算复杂度随输入图像尺寸平方增长，导致高计算代价。此外，ViT 在无大规模数据的场景下性能有限，训练过程对数据增强和预训练依赖较大 [1]。

### 2.3 ConvNeXt 的提出

ConvNeXt 在保留 CNN 局部特征提取优势的基础上，通过整合现代化设计元素提升了模型的全局建模能力。这些改进包括使用更深层次的网络结构、LayerNorm 替代 BatchNorm、引入深度可分离卷积等 [3]。主要分为宏观和微观改进，如宏观改进：改变计算资源比例和提取特征层、引入 ResNext 和倒瓶颈结构和改进训练方式和增大卷积核大小，微观改进：减少激活函数并将激活函数替代为 GeLu 等。ConvNeXt 的设计思路为传统 CNN 提供了现代化改造的范例。修改后的结构如图 1所示

| | output size | • ResNet-50 | • ConvNeXt-T | ○ Swin-T |
|---|---|---|---|---|
| stem | 56×56 | 7×7, 64, stride 2<br>3×3 max pool, stride 2 | 4×4, 96, stride 4 | 4×4, 96, stride 4 |
| res2 | 56×56 | [1×1, 64<br>3×3, 64<br>1×1, 256] × 3 | [d7×7, 96<br>1×1, 384<br>1×1, 96] × 3 | [1×1, 96×3<br>MSA, w7×7, H=3, rel. pos.<br>1×1, 96<br>[1×1, 384<br>1×1, 96]] × 2 |
| res3 | 28×28 | [1×1, 128<br>3×3, 128<br>1×1, 512] × 4 | [d7×7, 192<br>1×1, 768<br>1×1, 192] × 3 | [1×1, 192×3<br>MSA, w7×7, H=6, rel. pos.<br>1×1, 192<br>[1×1, 768<br>1×1, 192]] × 2 |
| res4 | 14×14 | [1×1, 256<br>3×3, 256<br>1×1, 1024] × 6 | [d7×7, 384<br>1×1, 1536<br>1×1, 384] × 9 | [1×1, 384×3<br>MSA, w7×7, H=12, rel. pos.<br>1×1, 384<br>[1×1, 1536<br>1×1, 384]] × 6 |
| res5 | 7×7 | [1×1, 512<br>3×3, 512<br>1×1, 2048] × 3 | [d7×7, 768<br>1×1, 3072<br>1×1, 768] × 3 | [1×1, 768×3<br>MSA, w7×7, H=24, rel. pos.<br>1×1, 768<br>[1×1, 3072<br>1×1, 768]] × 2 |
| FLOPs | | $4.1 \times 10^9$ | $4.5 \times 10^9$ | $4.5 \times 10^9$ |
| # params. | | $25.6 \times 10^6$ | $28.6 \times 10^6$ | $28.3 \times 10^6$ |

图 1. ConvNext 结构图

## 3 本文方法

### 3.1 本文方法概述

本文基于 ConvNeXt 网络的公开实现，复现了其在 Cifar-100 上的性能表现。图 2 展示了 ConvNeXt 的基本模块结构。如图 2所示：

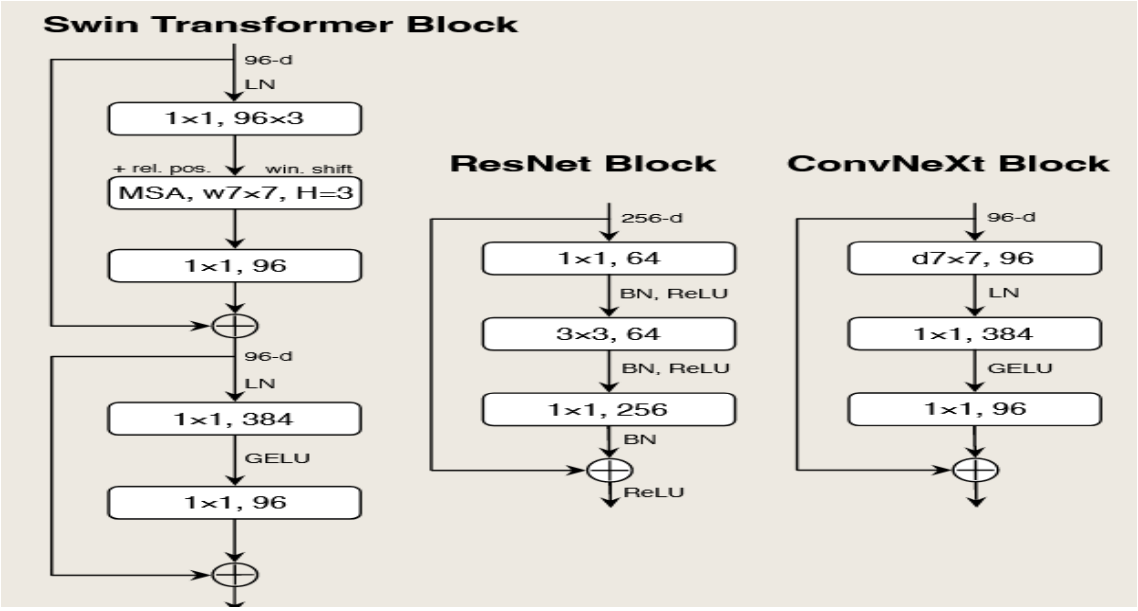


图 2. 方法示意图

### 3.2 特征提取模块

ConvNeXt 的特征提取模块通过深度可分离卷积实现了计算效率的提升，同时利用 LayerNorm 保证了数值稳定性。其模块主要包括：深度可分离卷积：减少参数量和计算量；GELU 激活函数：提升非线性表达能力；跨通道交互操作：通过 1×1 卷积实现特征融合。而在传统的 RestNet 中，stem cell 使用了 7×7 卷积层，步幅为 2，并随后进行最大池化导致输入图像在空间维度上缩小 4 倍。此处借鉴 Transformer 的分块策略，简化提取过程，改为 4×4 卷积层，步幅为 4。

### 3.3 损失函数定义

本文实验采用交叉熵损失函数（CrossEntropy Loss）作为分类任务的目标优化函数。此外，在多任务实验中结合了 Dice Loss 以适应语义分割任务的需求。
交叉熵损失（Cross-Entropy Loss）定义为：

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(p_i)$$

其中：

- $N$ 是类别的数量，
- $y_i$ 是第 $i$ 类的真实标签（通常采用独热编码），
- $p_i$ 是第 $i$ 类的预测概率，通常通过 softmax 函数得到。

对于多类别分类问题，损失可以对样本数量 $M$ 进行平均：

$$\mathcal{L} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N} y_i^{(m)} \log(p_i^{(m)})$$

其中：

- $M$ 是样本的数量，
- $y_i^{(m)}$ 是第 $m$ 个样本的第 $i$ 类真实标签，
- $p_i^{(m)}$ 是第 $m$ 个样本的第 $i$ 类预测概率。

# 4 复现细节

## 4.1 与已有开源代码对比

ConvNext 在 Github 上有开源项目，地址为ConvNext，在原代码基础上，修改了 batch-size 等超参数，添加了数据增强模块，简化了部分冗余模块，并试图添加 swin-mlp 模块提升 ConvNext 性能。但这些改进，并未在 CIFAR-100 上的产生明显效果。猜测数据加强已不能有效提升模型性能，应考虑提升 ConvNxet 捕获全局信息能力和加强局部依赖。

## 4.2 实验环境搭建

实验使用的是 pytorch 版本，其各版本信息为：pytorch==2.4.0；torchvision==0.19.0；torchaudio==2.4.0；pytorch-cuda==12.4，在 2 张 2080ti 上训练了 300 个 epoch，学习率初始值为 0.004，批量大小为 64, 运行大概 20 个小时。

## 4.3 创新点

ConvNext 融合了 Transformer 的设计思想，通过调整传统卷积神经网络（CNN）架构，使其更接近 Transformer 结构的特点，并使用更大卷积核来提高模型感受野。ConvNext 将卷积神经网络和 ViT 相结合，通过优化卷积层设计、激活函数的替换、使用更大感受野的卷积核、改进网络结构以及优化训练过程，成功提高了卷积网络的性能，同时保持了计算的高效性。通过这些创新，ConvNeXt 在多个视觉任务中实现了比传统 CNN 更好的表现，并且在硬件效率方面表现出色。

# 5 实验结果分析

在 CIFAR-100 数据集上验证了 ConvNeXt 的性能，并与 ResNet 和 ViT 进行了对比。

## 5.1 分类任务结果

CIFAR-100：ConvNeXt-Tiny 在 CIFAR-100 上 top-1 达到了 82.35%，与此相同参数量的 ViT 和 RestNet-1001 分别为 78.27% 和 77.3%。而进行改进的 ConvNext 与原 ConvNext 性能上差别不大，达到了 82.34%。

| 模型 | 参数量 (M) | Top-1 准确率 (%) | 推理速度 (ms/图像) |
|---|---|---|---|
| ResNet-1001 | 44.5 | 77.3 | 5.2 |
| ViT-B | 85.8 | 78.27 | 8.3 |
| ConvNeXt-T | 28.6 | 82.35 | 4.1 |
| ConvNeXt* | 28.6 | 82.34 | 4.1 |

表 1. ConvNext-ViT-RestNet 性能对比,* 代表改进后的 ConvNext

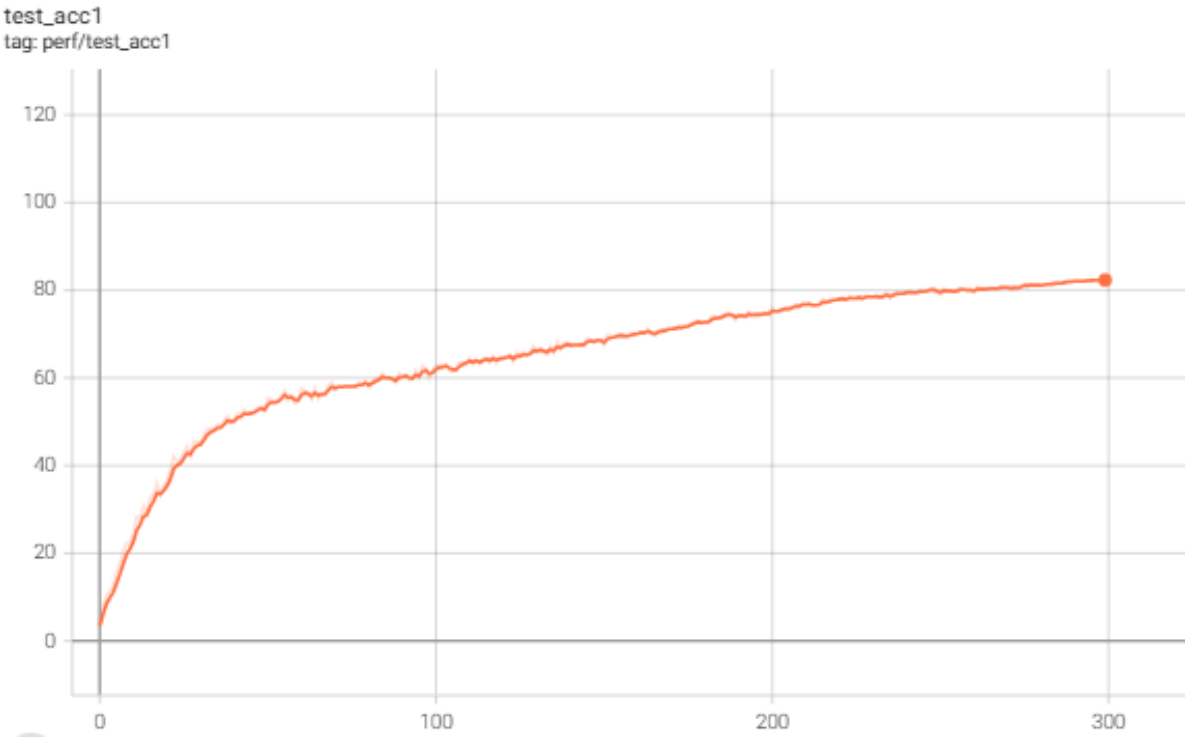


图 3. 实验结果 top-1 示意

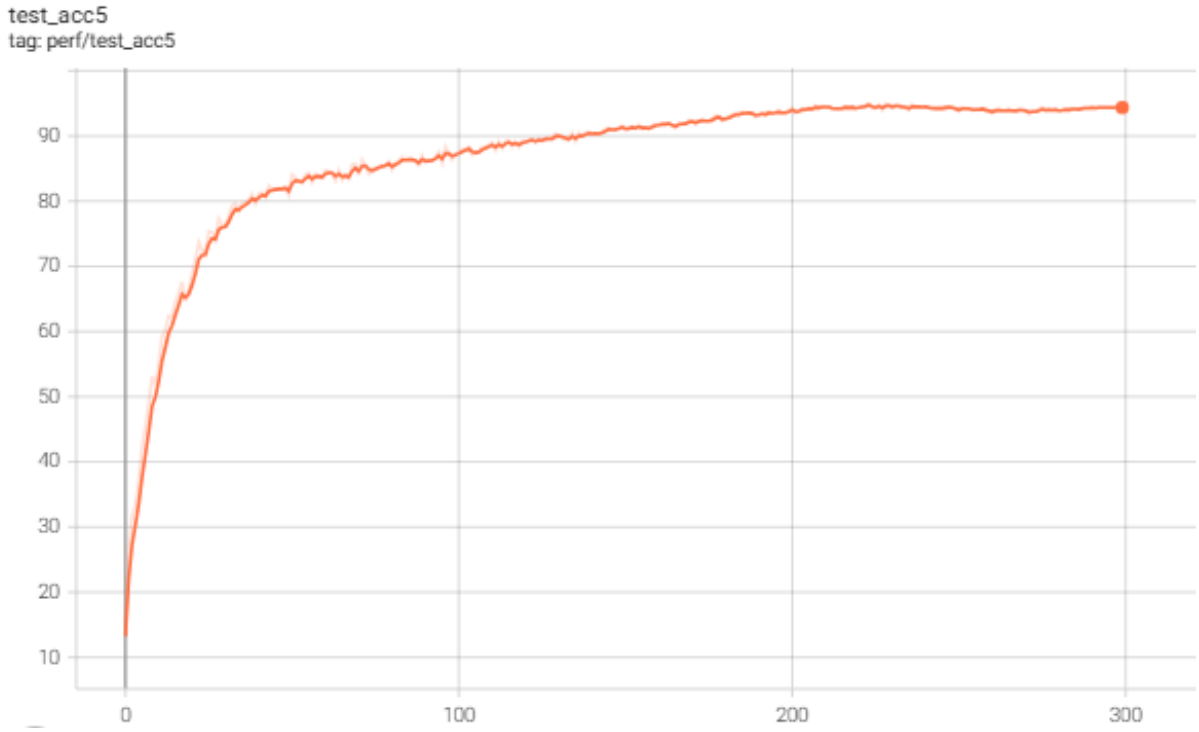


图 4. 实验结果 top-5 示意

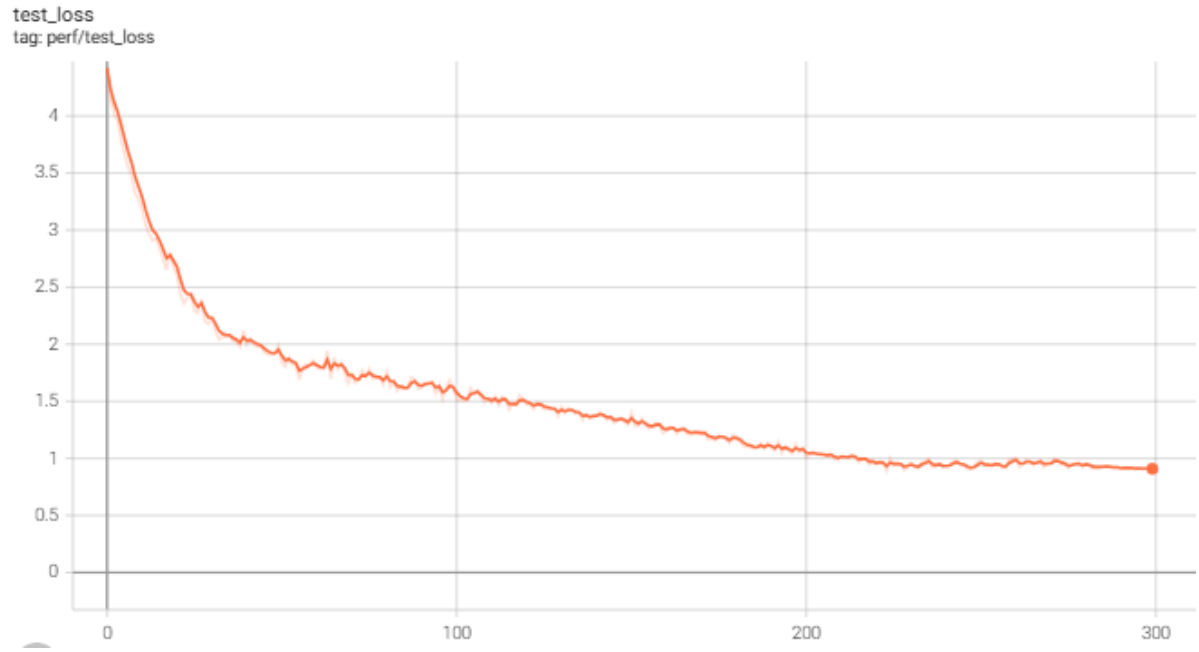


图 5. 实验结果 loss 示意

由此可知 ConvNext 在性能上已经媲美甚至超过以 Transformer 为 backbone 的部分视觉模型，其在视觉任务上的潜力巨大。上图的也展示了训练过程中 text-acc、loss 的变化。学习率由于使用了 AdamW 先上升后减少，保证训练初期更快收敛和后期更精细调参，并设置 min-lr 防止训练过于缓慢。loss 和 test-acc 的变化也说明模型训练较为稳定未出现发散情况。且虽然 ConvNext 卷积核大小为 7，但较于 Transformer 模型来说，全局信息的捕获能力较为缺失，应考虑提升模型的长距离依赖能力。

### 5.2 消融实验

为了验证 ConvNeXt 设计的有效性，我们进行了以下消融实验：LayerNorm 替代 BatchNorm：准确率提升约 1.5%；深度可分离卷积：减少了约 30% 的参数量，同时性能提升约 0.8%。如图 6所示。

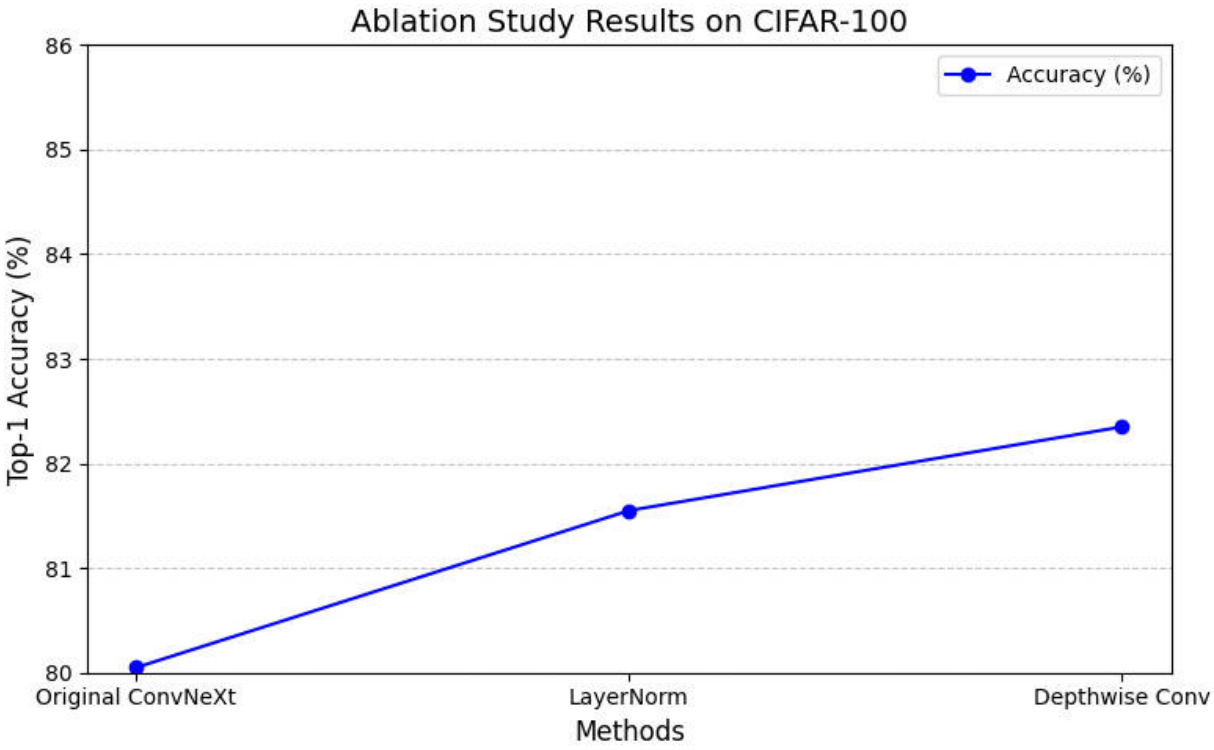


图 6. 消融实验

## 6 总结与展望

本文深入研究了 ConvNeXt 网络的设计思想与应用性能。通过复现实验，我们验证了其在 CIFAR-100 数据集上的优越性，同时分析了与 ViT 和 ResNet-101 的对比结果。实验表明，ConvNeXt 在参数量、精度和推理速度方面取得了良好的平衡。

未来的研究方向包括探索 ConvNeXt 在目标检测、语义分割等其他任务中的应用，结合神经架构搜索（NAS）进一步优化网络结构，以及在低算力设备上推广轻量化变体的适用性。

在 2020 年代，视觉 Transformers，特别是像 Swin Transformers 这样的层次结构，开始取代 ConvNet，成为通用视觉主干的首选。人们普遍认为，视觉 Transformer 比 ConvNet 更准确、更高效、更可扩展。而新提出的 ConvNeXts，是一种纯 ConvNet 模型，可以在多个计算机视觉基准上与最先进的分层视觉 Transformer 竞争，同时保留标准 ConvNet 的简单性和效率，对此卷积的特性在视觉任务中应重新重视和参考。而对于未来的研究中，也可以向两者相互融合的方向进行，如使用深度卷积代替 Swin 中的 MSA 块，或将 ConvNet 中的某些结构向 Transformer 更加靠近，充分利用各自优点，结合全局和局部注意力使视觉任务取得更大突破。也应多尝试以 ConvNext 为骨干框架的下游任务。

## 参考文献


[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022.