计算机科学
COMPUTER SCIENCE

面向图像数据的ConvNeXt特征提取研究

杨鹏跃, 王锋, 魏巍

引用本文

杨鹏跃, 王锋, 魏巍. 面向图像数据的ConvNeXt特征提取研究[J]. 计算机科学, 2024, 51(6A): 230500196-7.

YANG Pengyue, WANG Feng, WEI Wei. ConvNeXt Feature Extraction Study for Image Data[J]. Computer Science, 2024, 51(6A): 230500196-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500196

面向图像数据的 ConvNeXt 特征提取研究

杨鹏跃　王　锋　魏　巍
山西大学计算机与信息技术学院　太原 030006
(ypy1754660523@163.com)

摘　要　卷积神经网络在计算机视觉任务中已取得诸多成果,无论是目标检测还是分割,都依赖于提取到的特征信息,一些模糊性的数据和物体形状各异等问题为特征提取带来了极大的挑战。传统的卷积结构只能学习到特征图相邻空间位置的上下文信息,无法对全局信息进行提取,而自注意力机制等模型虽具有更大的感受野和建立全局的依赖关系,但存在计算复杂度过高和需要大量数据等不足。为此,提出了一种 CNN 与 LSTM 结合的模型,该模型在增强局部感受野的前提下,可以更好地结合图像数据的全局信息。研究以主干网络 ConvNeXt-T 为基础模型,通过拼接不同大小卷积核以融合多尺度特征来解决物体形状各异的问题,并从水平和垂直两个方向聚合双向长短期记忆网络关注全局与局部信息的交互性。实验对公开访问的 CIFAR-10,CIFAR-100,Tiny ImageNet 数据集进行图像分类任务,所提出的网络在 3 个数据集实验中相较于基础模型 ConvNeXt-T 在准确率上分别提高了 3.18%,2.91%,1.03%。实验证明改进后的 ConvNeXt-T 网络相较于基础模型在参数量和准确性方面都有了大幅度提升,可提取到更加有效的特征信息。

关键词:特征提取;局部感受野;ConvNeXt-T;多尺度特征;双向长短期记忆网络
中图分类号　TP391

ConvNeXt Feature Extraction Study for Image Data

YANG Pengyue, WANG Feng and WEI Wei
School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China

Abstract　Convolutional neural networks have achieved many results in computer vision tasks, both in target detection and segmentation, which depend on the extracted feature information. Some problems such as ambiguous data and varying object shapes pose great challenges for feature extraction. The traditional convolutional structure can only learn the contextual information of the neighboring spatial locations of the feature map and cannot extract the global information, while models such as the self-attentive mechanism, although having a larger perceptual field and establishing global dependencies, are insufficient due to their high computational complexity and the need for large amounts of data. Therefore, this paper proposes a model combining CNN and LSTM, which can better combine the global information of image data while enhancing the local perceptual field. It uses the backbone network ConvNeXt-T as the base model to solve the problem of different object shapes by splicing different size convolutional kernels to fuse multi-scale features, and aggregates two-way long and short-term memory networks from both horizontal and vertical directions. Focus on the interactivity of global and local information. Experiments are conducted on publicly accessible CIFAR-10, CIFAR-100, and Tiny ImageNet datasets for image classification tasks, and the accuracy of the proposed network improves 3.18%, 2.91%, and 1.03% in the three datasets respectively, compared to the base model ConvNeXt-T. Experiments demonstrate that the improved ConvNeXt-T network has substantially improved the number of parameters and accuracy compared with the base model, and can extract more effective feature information.

Keywords　Feature extraction, Local receptive field, ConvNeXt-T, Multi-scale features, Bidirectional long and short-term memory network

1　引言

卷积神经网络[1-3]允许将通用特征学习方法用于各种视觉识别任务[4-5],而不依赖于人工特征工程,这对计算机视觉领域产生了重大影响。2012 年,AlexNet[2] 的提出促成了“ImageNet 时刻”[6],开创了计算机视觉的新时代。自此之后,如 VGGNet[7],Inceptions[8],ResNe(X)t[9],DenseNet[10],MobileNet[11],EfficientNet[12] 和 RegNet[13] 等众多代表性卷积神经网络相继问世。随着模型参数量的激增,利用大规模预训练模型[14]进行微调已被广泛研究并用于视觉任务[15-16],常见方法包括微调部分参数和添加额外的残差块等[17-18]。在视觉任务中,一些模糊性的数据和物体形状各

基金项目:国家自然科学基金(62276158);山西省回国留学人员科研资助项目(2021-007)
This work was supported by the National Natural Science Fundation of China(62276158) and Research Project Supported by Shanxi Scholarship Council of China(2021-007).
通信作者:王锋(sxuwangfeng@126.com)

异等问题为特征提取带来了极大的挑战。自注意力机制等模型因其极佳的全局建模能力而在一些任务中表现突出，而本文提出了一种 CNN 与 LSTM 结合的方法，改进了传统 CNN 结构无法获得全局信息的弊端。本文的主要贡献包括 3 个方面：

1）基于 ConvNeXt-T 模型提出了新的改进策略，在无需大规模的预训练和强大的数据扩充情况下，在公开的数据集上取得了良好的结果；

2）分离出的下采样层，以较小的步幅保证了数据信息的不丢失，同时通过利用不同大小的卷积核在图像数据中提取不同的特征信息，增强了模型局部的感受野；

3）CNN 与长短期记忆网络的结合，在充分利用卷积神经网络本身内置的感应偏置学习到特征图相邻空间位置的上下文信息外，还增强了图像全局信息和局部信息的交互性。

2 相关工作

注意力机制[19-21]因在连续数据中权衡不同特征的能力而在 NLP 研究中广受欢迎。Transformer[22]作为一个完全基于注意力的模型引入，主要用于机器翻译和一般的 NLP。在此之后，基于注意力的模型，特别是变换器被应用于机器翻译之外的各种任务[23-25]，包括视觉问题回答[26-27]、动作识别[28-29]等。许多研究人员还利用神经网络中注意力和卷积的组合来完成视觉任务[30-33]。在计算机视觉领域，最近对端到端 Transformers[34-35]和 MLP[36-37]的兴趣激增，这促使人们用通用神经架构来学习输入数据的特征信息。Transformers 因注意力机制而表现突出，但也面临着一些挑战，如忽略了图像的 2D 结构，高分辨率图像的序列长度过长而导致计算复杂度过高。而 ConvNeXt[38]的提出，表明了纯卷积所带来的性能表现并不比 Transformer 差，甚至更佳。虽然更大的模型和数据集可以全面提高性能，但也带来了一系列挑战。Vision Transformer[34]，Swin Transformer[35]和 ConvNeXt 都以其巨大的模型变体获得了最佳表现，使得研究这些模型设计不可避免地导致了碳排放的增加。

相较于 CNN，RNN 因能够捕获长距离依赖而被广泛应用于文本相关任务中，但它在训练的过程中会出现梯度消失。而 LSTM 和门控循环单元网络引入门机制，较好地克服了 RNN 中梯度消失的弊端。Tang 等[39]以 CNN 或 LSTM 实现单句表示，而后用 gatedRNN 编码句子间的内在关系和语义联系，最终实现了文本的编码并较好地捕获了句子间的语义信息。Lai 等[40]提出了一种 RCNN 模型，首先利用双向循环网络模型得到单词的上下文信息，然后通过 CNN 的卷积池化过程进行文本分类，最终在 SST 等数据集上取得了最好的效果。

现实生活中，一些应用场景的图像数据并不清晰，甚至我们需要关注的某个人或者某个物体在整个图像中占比非常小，这给一些研究带来了极大的挑战。如 Tiny ImageNet[41]数据集因分辨率低，人类肉眼很难分辨出该图片信息。为此，我们首先对不同数据集采用了不同的预处理方法。Tiny ImageNet采用 Mixup[42]等数据增强方法，而 CIFAR-10 和 CIFAR-100[43]并未采用过多的数据增强方法。为了解决 CNN 无法获得全局信息以及大模型大数据对于计算资源的巨大需求等问题，本文以 ConvNeXt-T 为基础模型，充分结合 CNN 与 LSTM[44]的优势，旨在改进 CNN 的全局建模能力。

3 基于 ConvNeXt 的模型

本文选择 ConvNeXt-T 网络作为基础网络，与 Transformer 网络相比，ConvNeXt-T 网络不需要进行分块合并、窗口偏移和相对位置编码等操作，具有更好的性能和更少的计算量。ConvNeXt-T 网络的整体结构与 ResNet 相似，主干结构含 4 个计算个数为 3:3:9:3 的 Stage，每个 Stage 通道数为[96,192,384,768]，每个 Stage 层级间以 4 倍或 2 倍的下采样率进行特征抽取，逆瓶颈层（中间大、两头小）的基本块结构可有效避免信息流失。

本网络在 ConvNeXt-T 网络的基础上重新设计下采样和基本块，在保证每个 Stage 个数不变情况下，将每个 Stage 通道数缩减为[64,128,320,512]，层级式的网络结构可方便应用于检测等下游任务。而后，我们新加入一个双向 LSTM 层，通过特征图空间信息得到进一步重利用并传递到全连接层，最终经 Softmax 分类输出。LSTM 层考虑先前信息并关系到当前信息，增强全局信息与局部信息的交互性。本网络的总体框架结构如图 1 所示。

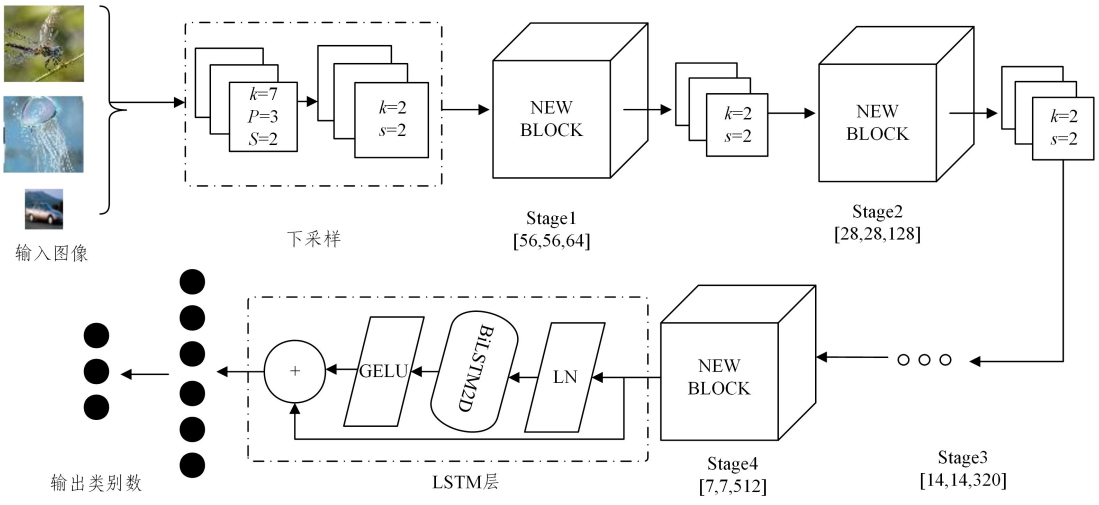


图 1 网络整体结构图

Fig. 1 Network overall structure diagram

3.1 下采样

在 ResNet 中，空间下采样是通过每个阶段开始时的残差块实现的。Swin Transformer 则在每个阶段之间添加了单独的下采样层，ConvNeXt 也是如此。在网络开始阶段，对输入图像做下采样操作，ResNet 是由卷积和最大池化层组成，Swin Transformer 以及 ConvNeXt 则是由大小为 4、步幅为 4 的卷积组成。由 Sunkara 等提出的 SPD-Conv[45]，在图像分辨率较低或物体较小的情况下，用池化层或者步幅大于 1 的卷积，会导致细粒度信息的丢失和对低效特征表示的学习。Swin Transformer 中采用大小为 4、步幅为 4 的卷积是为了方便计算而分割成不重叠的窗口。这里我们采用大小为 7 和 2、步幅为 2 的卷积。

表 1 不同模型的下采样方法

Table 1 Downsampling methods for different models

ResNet-50	ConvNeXt-T	Our ConvNeXt-T
7×7，s=2，64 3×3 Maxpool，s=2	4×4，s=4，96	7×7，s=2，64 2×2，s=2，64

3.2 LSTM 层

1) BiLSTM

LSTM 是具有记忆长短期信息的能力的神经网络，它通过引入遗忘门、输入门、输出门机制用于控制特征的流通和损失，有效解决了 RNN 的长期依赖问题。基本结构如图 2 所示。

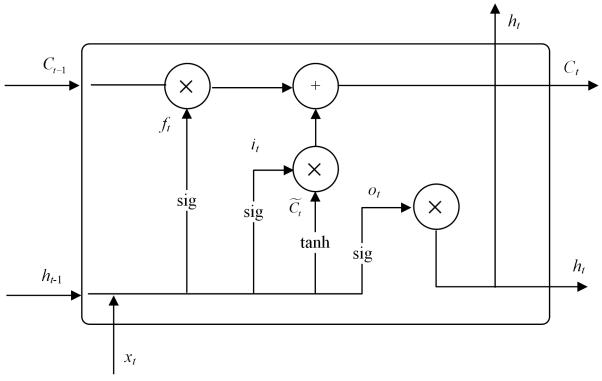


图 2 LSTM 基本结构图

Fig. 2 Basic structure of LSTM

图中，C_t 表示单元状态，C_{t-1} 表示上一时刻的单元状态，h_t 表示输出信号，h_{t-1} 表示隐藏层状态为上一时刻的输出信号。计算公式如下：

$$f_t=\mathrm{Sigmoid}(W_f\cdot[h_{t-1},x_t]+b_f) \tag{1}$$

$$i_t=\mathrm{Sigmoid}(W_i\cdot[h_{t-1},x_t]+b_i) \tag{2}$$

$$\widetilde{C}_t=\tanh(W_C\cdot[h_{t-1},x_t]+b_C) \tag{3}$$

$$C_t=f_t*C_{t-1}+i_t*\widetilde{C}_t \tag{4}$$

$$o_t=\mathrm{Sigmoid}(W_o\cdot[h_{t-1},x_t]+b_o) \tag{5}$$

$$h_t=o_t*\tanh(C_t) \tag{6}$$

其中，式(1)表示遗忘门，决定什么信息需要从单元状态中删除；式(2)－式(4)表示更新门，决定哪些新的信息需要增加至单元状态中；通过计算式(5)、式(6)，得到输出值。

BiLSTM 由两个普通的 LSTM 组成，一个是正向去处理输入序列，另一个反向处理序列，处理完成后将两个 LSTM 的输出拼接起来。具体公式如式(7)－式(9)所示：

$$\overrightarrow{h_{\mathrm{for}}}=LSTM_{\mathrm{for}}(\vec{x}) \tag{7}$$

$$\overleftarrow{h_{\mathrm{back}}}=LSTM_{\mathrm{back}}(\overleftarrow{x}) \tag{8}$$

$$h=concatenate(\overrightarrow{h_{\mathrm{for}}},\overleftarrow{h_{\mathrm{back}}}) \tag{9}$$

其中，$\vec{x}$ 为输入序列，$\overleftarrow{x}$ 与 $\vec{x}$ 顺序相反，$\overrightarrow{h_{\mathrm{for}}}$ 是对正向序列的输出结果，$\overleftarrow{h_{\mathrm{back}}}$ 是对反向序列的输出结果。

2) BiLSTM2D 层

BiLSTM2D 层是一种有效混合 2D 空间信息的技术。它由两个普通的 BiLSTM 组成：一个垂直 BiLSTM 和一个水平 BiLSTM。式(10)计算垂直方向，式(11)计算水平方向。

$$H_w^{\mathrm{ver}}=BiLSTM(X_w) \tag{10}$$

$$H_h^{\mathrm{hor}}=BiLSTM(X_h) \tag{11}$$

其中，X_h 和 X_w 代表对输入特征 X 高度和宽度的序列。我们将得到的水平和垂直输出进行维度拼接，然后经过全连接进行融合。特征 X 最终的输出结果如式(12)、式(13)所示：

$$H=concatenate(H^{\mathrm{ver}},H^{\mathrm{hor}}) \tag{12}$$

$$X=FC(H) \tag{13}$$

其中，$FC(\cdot)$代表权重为 $W\in R^{C\times 4D}$ 的全连接，C 代表通道维度，D 代表隐藏维度。我们通过 LSTM 对输入特征水平和垂直维度进行建模，使模型增强全局建模能力。基本结构图如图 3 所示。

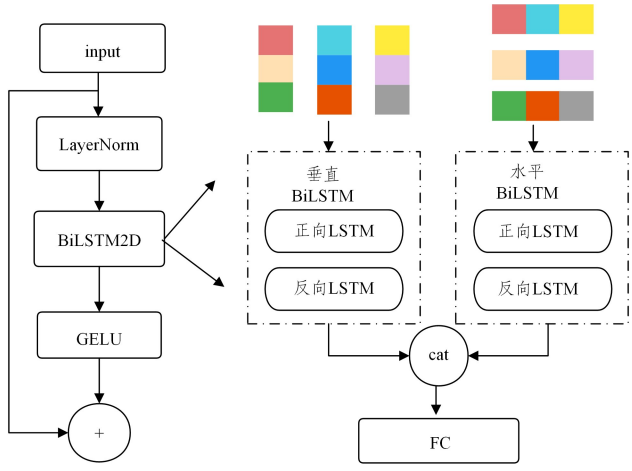


图 3 LSTM 层结构图

Fig. 3 Basic structure of LSTM

3.3 New Block

针对每个基本块的设计仍然保持“中间大，两头小”的网络结构，该想法由 MobileNetV2[46] 推广，与 ResNet 残差结构先降维，再升维(两头大，中间小)结构相反。受 GooLeNet[47] 启发，Inception 架构的主要想法是考虑怎样近似卷积视觉网络的最优稀疏结构并用容易获得的密集组件进行逼近和覆盖，我们在进行了大核卷积之后，不再简单使用单个 1×1 卷积进行维度的提升，而是采用多种卷积核大小进行特征提取。通过不同大小的滤波器提取特征之后，在通道维度上进行维度拼接，而后经过 1×1 卷积进行降维并通道融合。相较于 GooLeNet 中的 Inception 架构，我们不使用最大化池化操作。众所周知，无论是加深网络还是加宽网络，都会造成参数量和计算资源的增加。为了防止模型“过大”，在进行 3×3，5×5 卷积的前一个单元层，先使用 1×1 卷积以 2 倍比率进行降维。与 Inception 架构的另一个不同点则是，我们将最开始的 7×7 滤波器结果直接与下一层的特征图拼接，一方面进行特征复用，降低了参数量，另一方面，通过多个卷积核提取图像不同尺度的信息，最后进行融合，可以得到图像更好的表征。原始的 ConvNeXt 和我们改进后的 ConvNeXt 如图 4 所示。

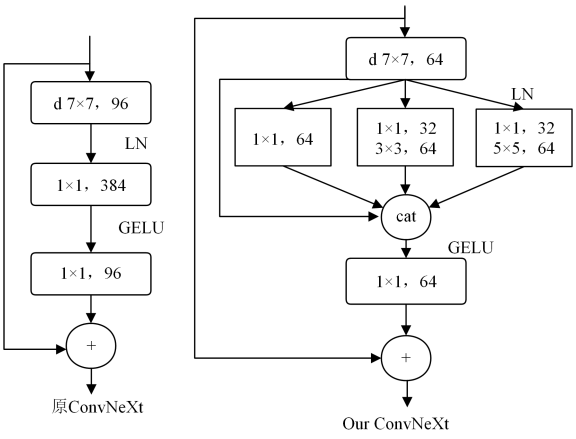


图 4 New Block 结构图

Fig. 4 Diagram of New Block structure

4 实验数据及实验设置

本节主要介绍常用的公开的实验数据集，并进行相应的数据预处理和实验设置的相关说明。

4.1 实验数据

Tiny ImageNet：它是拥有 120000 张标签图像的数据集，由 200 个对象类别组成。这些类别是 WordNet 层次结构的集合，图像在本质上与 ILSVRC 基准中使用的 ImageNet[48] 图像相似，但分辨率较低。所有这些图像都下采样到 64×64 的固定分辨率。通过从每个图像中减去整个 Tiny ImageNet 数据集的平均图像来对所有图像进行预处理，因此模型的输入是每个图像的平均中心 RGB 像素值。从人类肉眼来看，有些图像很难分类。

CIFAR-10/100：CIFAR-10 由 10 个类的 60000 个 32×32 彩色图像组成，每个类有 6000 个图像。CIFAR-100 包含 100 个类，每个类有 500 个训练图像和 100 个测试图像。具体信息如表 2 所列。

表 2 不同数据集数量和类别信息

Table 2 Number and category information of different datasets

数据集	分辨率	训练集	验证集	测试集	类别
Tiny ImageNet	64×64	100000	10000	10000	200
CIFAR10	32×32	50000	—	10000	10
CIFAR100	32×32	50000	—	10000	100

4.2 数据预处理

对于小型数据集来说，过拟合问题是关键性问题，通过数据增强，可以让有限的数据产生等价于更多数据的价值，进一步使模型学习到更多鲁棒性的特征，从而有效提高模型的泛化能力。我们使用 Mixup 将两张图片直接进行线性组合，对应的，标签也进行线性组合。其中，叠加在背景图片上的图片强度按照 $\beta(\alpha,\alpha)$ 分布随机采样获得，α 为超参数，这里设置为 0.2。在训练时我们还引入了随机裁剪、50%概率随机水平翻转和随机擦除等方法对图像进行特征增强。为了图片清晰(Tiny 分辨率低)，图 5 给出了 ImageNet 增强后的数据集示例。

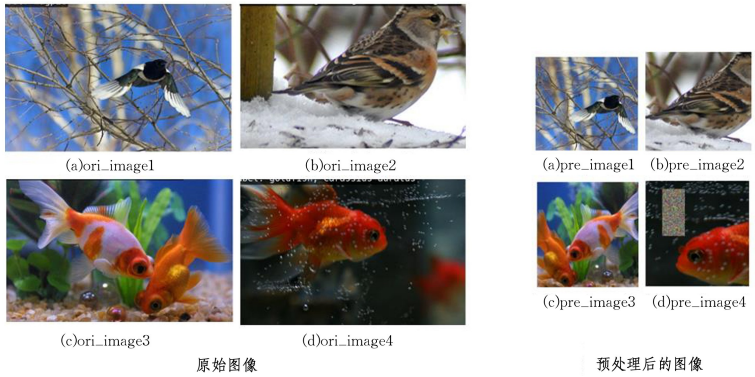


图 5 数据增强过程

Fig. 5 Data enhancement process

4.3 实验设置

Tiny ImageNet：考虑到计算资源的使用和消耗问题，我们并没有进行更大模型变体的实验，完全基于上面所设计的网络规模，另外对于一些超参数(如学习率等)的设置并没有进行更深入的研究。在 Tiny ImageNet 数据集上，为防止随机性对实验的影响，我们自始至终保持随机种子为 3407，从头开始训练数据集，并不借助训练权重。本实验基于开源框架 OpenMMLab 进行，使用 Adam 优化策略并以 4×10^{-3} 的学习率训练模型 300 个 epoch。

CIFAR10/100：为证明改进后的 ConvNeXt 在图像分类任务上的实用性，我们在图像数量和图像大小的多种类型的公开可用(根据 MIT 许可证)数据集上也进行了实验。数据集包括 CIFAR-10 和 CIFAR-100。实验中我们使用学习率为 0.1 的随机梯度下降策略训练 CIFAR 数据集 200epoch，详细参数如表 3 所列。

表 3 实验详细参数设置

Table 3 Detailed experimental parameter settings

	Tiny ImageNet	CIFAR-10/100
optimizer	AdamW	SGD
base learning rate	4×10^{-3}	0.1
weight decay	0.05	0.0001
optimizer momentum	$\beta_1,\beta_2=0.9,0.999$	0.9
batch size	256	128
training epochs	300	200
learning rate schedule	cosine decay	Step
warmup epochs/step	20	[100,150]
warmup schedule	linear	None
RandomCrop	224	32
RandomFlip	0.5	0.5
randaugment	(9,0.5)	None
mixup	0.2	None
random erasing	0.25	None
label smoothing	0.1	None
layer scale	1×10^{-6}	None

5 实验结果及分析

本节通过进行相应的实验，论证本文的改进方法是可行的。我们通过与其他实验数据的对比和消融实验，直观地感受改进后模型的表现能力。

5.1 对比实验

为了验证本文的改进方法，我们选取了一些基线网络，通过准确率、参数量等评价指标进行对比。从表 4 可以看出，改进后的 ConvNeXt 相比其他的基线模型，表现出了优异的结果，且相比于原始的 ConvNeXt 也有一定的突出优势。

表 4 各模型在 Tiny ImageNet 数据集上的性能对比

Table 4 Performance comparison of each model on Tiny ImageNet dataset

Model	Accuracy/%	#Param/GFLOPs	GPU RAM for batch size 64/GB
ResNet-18 (Hassani et al. 2021)	48.11	11.20×10^6	1.2
ResNet-34 (Hassani et al. 2021)	45.60	21.30×10^6	1.4
ResNet-50 (Hassani et al. 2021)	48.77	25.20×10^6	3.3
ViT-128/4×4	26.43	0.53×10^6	13.8
CCT-128/4×4	39.05	0.91×10^6	15.8
CvT-128/4×4	40.69	1.12×10^6	15.4
Wavemix-Lite-160/13	54.76	6.90×10^6	9.4
WaveMix-Lite-256/7	51.37	9.62×10^6	2.3
ResNeXt-50 (AutoMix)	70.72	24.84×10^6	—
ConvNeXt-T*	**70.36**	**27.97×10^6/4.46**	**9.54**
Our ConvNeXt-T*	**71.39**	**12.75×10^6/2.19**	**8.46**

注：该表列出了不同模型对于 Tiny ImageNet 的准确率、参数量等指标，以输入尺寸为 224 计算 GFLOPs 数值。表中带 * 表示自己实验跑出的数据，而不带 * 表示直接引用论文中的结果。

对于 CIFAR-10 和 CIFAR-100，从表 5 的统计结果可以看出，本文的改进方法性能显著提高了 3.18%，2.91%。随着我们进行超参数的选取和模型对数据集的适应性操作，当以学习率为 1.25×10^{-3}、权重衰减为 0.05、优化器选择 AdamW 进行模型的训练，所提方法的准确率相较于基线模型也有出色的表现能力。

表 5 各模型在 CIFAR10 和 CIFAR100 数据集上的性能对比

Table 5 Performance comparison of various models on CIFAR10 and CIFAR100 datasets

Model	Accuracy/%		#Param.	MACs
	CIFAR-10	CIFAR-100		
ResNet-18	90.27	66.46	11.18×10^6	0.04×10^9
ResNet-34	90.51	66.84	21.29×10^6	0.08×10^9
ResNet56[49]	94.63	74.81	0.85×10^6	0.13×10^9
ResNet110[49]	95.08	76.63	1.73×10^6	0.26×10^9
CVT-7/8	89.79	70.11	3.74×10^6	0.06×10^9
ViT-12/16	83.04	57.97	85.63×10^6	0.43×10^9
ViT-Lite-7/16	78.45	52.87	3.89×10^6	0.02×10^9
ViT-Lite-7/8	89.10	67.27	3.74×10^6	0.06×10^9
ConvNeXt-T*	**79.26**	**52.45**	**27.83×10^6**	**0.09×10^9**
Our ConvNeXt-T*	**82.44**	**55.36**	**12.65×10^6**	**0.05×10^9**
Our ConvNeXt-T	**92.70**	**72.76**	**12.65×10^6**	**0.05×10^9**

5.2 消融实验

本实验基于 CIFAR10 数据集进行消融实验，固定任何超参数的设置，只针对网络结构进行实验分析，保证数据的真实性和网络结构的可行性。

在保证通道数为[64，128，320，512]，以及 3∶3∶9∶3 条件一致的情况下，对于新改进的基本块进行了实验数据分析，实验结果如表 6 所列。我们在原有的 ConvNeXt 模型基础上进行实验，相应的实验构建如下：

模型 A：将原有的基本块替换为改进后的 new block；模型 B：原有模型的基础上只添加 LSTM 层；模型 C：移除原先的 $k=4$，$s=4$ 下采样，添加改进后的下采样；模型 D：改进后的下采样加改进后的 new block；模型 E：下采样不变，改变后的 new block 加 LSTM 层；模型 F：改变后的下采样加 LSTM 层，block 不变。

表 6 不同模型在 CIFAR10 数据集上的消融实验对比

Table 6 Comparison of ablation experiments of different models on CIFAR10 dataset

	Accuracy/%		#Params	GFLOPs
	Top-1	Top-5		Shape=224
ConvNeXt-T/512	80.52	98.89	15.31×10^6	2.55
模型 A	80.05	98.92	11.05×10^6	2.01
模型 B	79.23	98.81	16.89×10^6	2.56
模型 C	80.38	98.78	15.33×10^6	2.71
模型 D	81.37	99.08	11.08×10^6	2.17
模型 E	81.23	99.19	12.63×10^6	2.03
模型 F	76.42	98.54	16.91×10^6	2.72
Our ConvNeXt-T	82.44	99.08	12.65×10^6	2.19

注：ConvNeXt-T/512 表示模型结构为原始的 ConvNeXt-T，只是通道数为[64，128，320，512]，这样做是为了防止通道数对实验结果的影响。

表 6 的实验结果表明：对原 ConvNeXt-T 只进行单个模块单元的改变，都会造成精度的下降。其中，只添加 LSTM 下降了 1%左右；若将原有的基本块替换成 new block，则无论添加其他任何一个，都会提升精度。本文模型首先通过较小的步幅进行下采样，以减少特征信息的丢失；然后，通过不同大小的卷积核进行多尺度信息的提取；最后，通过 LSTM 进行全局信息和局部感受野的信息融合，以达到优异的实验结果。

在进行了定量的消融实验后，我们对 new block 进行了新的研究，表 7 列出的数据进一步证实了本文模型具有良好的表现能力。针对不同的设计主要体现在增加新的卷积核大小(7×7)、两个 1×1 卷积以及 3×3 卷积换成 1×3 和 3×1 的结合。不同的设计结构图如图 6 所示。

表 7 不同 new block 设计在 CIFAR10 和 Tiny ImageNet 上的实验结果

Table 7 Experimental results of different new block designs on CIFAR10 and Tiny ImageNet

	CIFAR10-Accuracy/%		Accuracy/%	#Params	GFLOPs
	Top-1	Top-5	Tiny ImageNet		Shape=224
Our ConvNeXt	82.44	99.08	71.39	12.65×10^6	2.19
设计 A	75.86	98.23	70.02	12.02×10^6	1.97
设计 B	81.18	99.01	69.16	11.77×10^6	2.19
设计 C	80.16	98.86	69.40	11.69×10^6	1.86

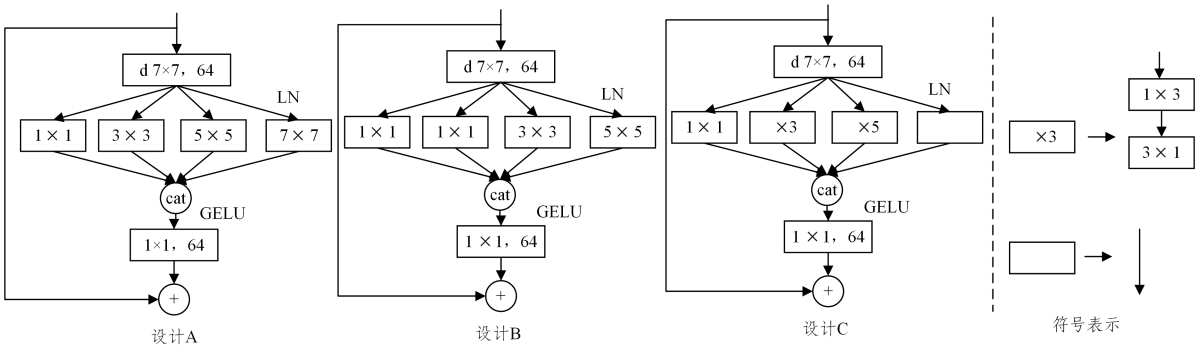


图 6 不同的 New Block 设计图

Fig. 6 Different new block designs

结束语 本文以 ConvNeXt-T 为出发点，为改进 ConvNeXt 在图像数据集上的表现，重新设计了下采样，并进行模型基本块的改进和整体模型设计。本文主要通过不同大小的滤波器学习到不同的感受野，并通过 LSTM 增强特征的全局信息。通过相应的实验证明，改进后的 ConvNeXt-T 在图像数据集上的准确率和参数量都有了明显的提升。出于计算资源和训练时间的考虑，本文只选择了在分辨率和数量较小的数据集上进行实验，未来我们将在 ImageNet 1K 或者更大的 ImageNet 22K 上证明所提模型具有相同出色的学习能力，同时通过降低卷积核大小以及调整下采样步幅进一步提高在 CIFAR-10 和 CIFAR-100 等数据集上的表现。另外，也可考虑将大规模预训练模型应用于各种计算机视觉的下游任务。

参 考 文 献

[1] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1026-1034.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[3] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.

[5] HE K M, GKIOXARI G, DOLLAR P, et al. Mask r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.

[6] RUSSAKOVSKY O, JIA D, HAO S, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115: 211-252.

[7] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. arXiv: 1406.2199, 2014.

[8] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[9] XIE S N, GIRSHICK R, DOLLAR 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.

[10] HUAN G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.

[11] HOWARD A G, ZHU M L, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv: 1704.04861, 2017.

[12] TAN M X, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]// International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[13] 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.

[14] WANG W H, BAO H B, DONG L, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks[J]. arXiv: 2208.10442, 2022.

[15] JIA M L, TANG L M, CHEN B C, et al. Visual prompt tuning[C]// 17th European Conference Computer Vision (ECCV 2022). Tel Aviv, Israel, Part XXXIII. Cham: Springer Nature Switzerland, 2022: 709-727.

[16] BAHNG H, JAHANIAN A, SANKARANARAYANAN S, et al. Visual prompting: Modifying pixel space to adapt pre-trained models[J]. arXiv: 2203.17274, 2022.

[17] JIA M L, WU Z X, REITER A, et al. Exploring visual engagement signals for representation learning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 4206-4217.

[18] YANG T J N, ZHU Y, XIE Y S, et al. Aim: Adapting image models for efficient video action recognition[J]. arXiv: 2302.03024, 2023.

[19] GRAVES A, WAYNE G, DANIHELKA I. Neural turing machines[J]. arXiv: 1410.5401, 2014.

[20] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409.0473, 2014.

[21] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv: 1508.04025, 2015.

[22] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. arXiv：1706.03762，2017.

[23] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv：1810.04805，2018.

[24] LIU Y，OTT M，GOYAL N，et al. Roberta：A robustly optimized bert pretraining approach[J]. arXiv：1907.11692，2019.

[25] YANG Z L，DAI Z H，YANG Y M，et al. Xlnet：Generalized autoregressive pretraining for language understanding[J]. arXiv：1906，08237，2019.

[26] LU J，BATRA D，PARIKH D，et al. Vilbert：Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. arXiv：1908.02265，2019.

[27] SU W，ZHU X，CAO Y，et al. Vl-bert：Pre-training of generic visual-linguistic representations[J]. arXiv：1908.08530，2019.

[28] BERTASIUS G，WANG H，TORRESANI L. Is space-time attention all you need for video understanding？[C]// ICML. 2021.

[29] GIRDHAR R，CARREIRA J，DOERSCH C，et al. Video action transformer network[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：244-253.

[30] WANG F，JIANG M Q，QIAN C，et al. Residual attention network for image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：3156-3164.

[31] HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：7132-7141.

[32] BELLO I，ZOPH B，VASWANI A，et al. Attention augmented convolutional networks[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：3286-3295.

[33] ZHANG H，GOODFELLOW I，METAXAS D，et al. Self-attention generative adversarial networks[C]// International Conference on Machine Learning. PMLR，2019：7354-7363.

[34] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：Transformers for image recognition at scale[J]. arXiv：2010.11929，2020.

[35] LIU Z，LIN Y T，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.

[36] TOLSTIKHIN I O，HOULSBY N，KOLESNIKOV A，et al. Mlp-mixer：An all-mlp architecture for vision[J]. Advances in Neural Information Processing Systems，2021，34：24261-24272.

[37] LIU H X，DAI Z H，SO D，et al. Pay attention to mlps[J]. Advances in Neural Information Processing Systems，2021，34：9204-9215.

[38] LIU Z，MAO H Z，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.

[39] TANG D Y，QIN B，LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]// Proceedings of the 2015 Conference on Empirical Methods in natural Language Processing. 2015：1422-1432.

[40] LAI S W，XU L H，LIU K，et al. Recurrent convolutional neural networks for text classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2015.

[41] LE Y，YANG X. Tiny imagenet visual recognition challenge[J]. CS 231N，2015，7(7)：3.

[42] ZHANG H Y，CISSE M，DAUPHIN Y N，et al. mixup：Beyond empirical risk minimization[J]. arXiv：1710.09412，2017.

[43] KRIZHEVSKY A，HINTON G. Learning multiple layers of features from tiny images[J]. https：//www. cs. toronto. edu/~kriz/learning-features-2009-TR. pdf.

[44] GRAVES A. Long short-term memory [J]. Supervised Sequence Labelling with Recurrent Neural Networks，2012：37-45.

[45] SUNKARA R，LUO T. No more strided convolutions or pooling：a new CNN building block for low-resolution images and small objects[C]// Machine Learning and Knowledge Discovery in Databases：European Conference（ECML PKDD 2022）. Grenoble，France，Part III. Cham：Springer Nature Switzerland，2023：443-459.

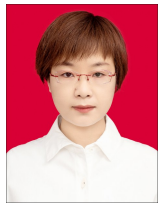
[46] SANDLER M，HOWARD A，ZHU M，et al. Mobilenetv2：Inverted residuals and linear bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：4510-4520.

[47] SZEGEDY C，VANHOUCKE V，IOFFE S，et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：2818-2826.

[48] DENG J，DONG W，SOCHER R，et al. Imagenet：A large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2009：248-255.

[49] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：770-778.

YANG Pengyue，born in 1999，postgraduate. His main research interests include image processing，data mining，and machine learning.



WANG Feng，born in 1984，Ph.D，is a member of CCF（No. 36494M）. Her main research interests include data mining，machine learning，and granular computing.