计算机科学
COMPUTER SCIENCE

LNG-Transformer:基于多尺度信息交互的图像分类网络

王文杰, 杨燕, 敬丽丽, 王杰, 刘言

引用本文

王文杰, 杨燕, 敬丽丽, 王杰, 刘言. LNG-Transformer:基于多尺度信息交互的图像分类网络[J]. 计算机科学, 2024, 51(2): 189-195.

WANG Wenjie, YANG Yan, JING Lili, WANG Jie, LIU Yan. LNG-Transformer:An Image Classification Network Based on Multi-scale Information Interaction [J]. Computer Science, 2024, 51(2): 189-195.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100218

LNG-Transformer:基于多尺度信息交互的图像分类网络

王文杰　杨　燕　敬丽丽　王　杰　刘　言
西南交通大学计算机与人工智能学院 成都 611756
可持续城市交通智能化教育部工程研究中心 成都 611756
(yyang@swjtu.edu.cn)

摘　要　鉴于Transformer的Self-Attention机制具有优秀的表征能力,许多研究者提出了基于Self-Attention机制的图像处理模型,并取得了巨大成功。然而,基于Self-Attention的传统图像分类网络无法兼顾全局信息和计算复杂度,限制了Self-Attention的广泛应用。文中提出了一种有效的、可扩展的注意力模块Local Neighbor Global Self-Attention(LNG-SA),该模块在任意时期都能进行局部信息、邻居信息和全局信息的交互。通过重复级联LNG-SA模块,设计了一个全新的网络,称为LNG-Transformer。该网络整体采用层次化结构,具有优秀的灵活性,其计算复杂度与图像分辨率呈线性关系。LNG-SA模块的特性使得LNG-Transformer即使在早期的高分辨率阶段,也可以进行局部信息、邻居信息和全局信息的交互,从而带来更高的效率、更强的学习能力。实验结果表明,LNG-Transformer在图像分类任务中具有良好的性能。

关键词:图像分类;自注意力机制;多尺度;Transformer
中图分类号　TP301

LNG-Transformer:An Image Classification Network Based on Multi-scale Information Interaction

WANG Wenjie,YANG Yan,JING Lili,WANG Jie and LIU Yan
School of Computing and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China
Engineering Research Center of Sustainable Urban Intelligent Transportation,Ministry of Education,Chengdu 611756,China

Abstract　Due to the superior representation capability of the Transformer's Self-Attention mechanism,several researchers have developed Self-Attention mechanism-based image processing model and achieved great success. However,the traditional network for image classification based on Self-Attention cannot take into account global information and computational complexity,which limits the wide application of Self-Attention. This paper proposes an efficient and scalable attention module,Local Neighbor Global Self-Attention(LNG-SA),that may interact with local,neighbor,and global information at any stage. By cascading LNG-SA module,a brand-new network called LNG-Transformer is created. LNG-Transformer adopts a hierarchical structure that provides excellent flexibility,and has a computational complexity proportional to image resolution. The features of LNG-SA enable LNG-Transformer to interact with local information,neighbor information,and global information even in the early stage of high-resolution,resulting in increased efficiency and enhanced learning capacity. Experimental results show that LNG-Transformer performs well at image classification.

Keywords　Image classification,Self-Attention,Multiple scales,Transformer

1　引言

在图像分类领域,卷积神经网络(CNN)一直占据主导地位。自AlexNet[1]在ImageNet图像分类比赛中取得第一名以来,大量学者投入到CNN的研究之中,他们使用更大的卷积核[2]、多分枝结构[3]、残差结构[4]、密集链接[5]、深度可分离卷积[6]、膨胀卷积[7]、参数重结构化[8]、注意力机制[9-11]、对比自监督学习[12]、融合标签间强相关性[13]等,在视觉领域的各个任务上都取得了不错的效果,进一步巩固了CNN在整个视觉领域的地位。

Transformer[9]是为文本序列任务而设计的,并且有效地缓解了梯度爆炸和梯度消失问题。之后由Transformer演化出来的Bert[14],更是成为了自然语言处理的主要网络。一些学者发现Transformer强大的拟合能力后,尝试将Transformer应用在图像领域,Vision Transformer[15]的发布正式开启了Trans-former在图像任务上的研究。

Vision Transformer网络的输入是把图像拆分成给定大小的窗口,加上位置信息后,直接送入网络中进行训练。虽然

到稿日期:2022-11-25　返修日期:2023-03-27
基金项目:国家自然科学基金(61976247)
This work was supported by the National Natural Science Foundation of China(61976247).
通信作者:杨燕(yyang@swjtu.edu.cn)

Vision Transformer 的每一层都能很好地学习到全局信息，但其每一层都需要大量参数拟合特征，导致必须以海量的训练数据为基础，才能发挥它的性能，给研究者们带来了很大的挑战。

随后发布的 Swin Transformer[16] 提出层次化结构和滑动窗口，很好地解决了 Vision Transformer 参数量大、难以训练的问题。通过把图像划分成一个个大小相同的窗口，只在窗口内进行信息交互，大幅度降低了计算复杂度，与输入图像大小呈线性关系；并且通过使用滑动窗口的方式，很好地学习到了邻居窗口信息。然而，窗口化带来了另一个问题，Swin Transformer 不能像 Vision Transformer 一样，在每一层都学到全部信息，只能通过增加网络深度，逐渐合并窗口之间的信息来学习全局信息。尽管与 Vision Transformer 中使用的全局信息交换相比，基于窗口内的信息交互具有更大的灵活性和泛化性，但网络在早期却丢失了全局信息，使得网络的学习能力有限。若在早期进行全局信息交换训练，则会导致计算量很大。因此，如何同时学习局部信息和全局信息成为了本文研究的重点。

为了解决上述问题，本文基于 Swin Transformer 提出了一个新的注意力模块 LNG-SA，以及 LNG-SA 模块的简化版本 Local Local Self-Attention(LL-SA)。LNG-SA 模块包含了局部信息、邻居信息和全局信息的交互，LL-SA 模块只包含两次局部信息交互。与 Viosn Transformer 和 Swin Transformer 相比，LNG-SA 具有更好的灵活性和更高的效率，在任何一个时期都可以学习到全局信息。

为了证明 LNG-SA 模块和 LL-SA 模块的有效性，我们设计了一个仅由多个简单的 LNG-SA 模块和 LL-SA 模块组成的视觉网络，命名为 LNG-Transforer，网络的主干部分没有使用任何卷积操作。大量实验表明，在保证近似的计算量和参数规模的情况下，LNG-Transformer 的性能优于 Swin Transfor-mer 和 ResNet[17]。在保证较少计算量和较小参数规模的情况下，LNG-Transformer 的性能大幅度领先于 Vision Transformer。

本文工作的主要贡献如下：

1）提出了一个简单有效的视觉网络——LNG-Transformer，该网络由多个简单的 LNG-SA 模块和 LL-SA 模块级联而成，在早期或者任何一个时期都可以学习到全局信息。

2）LNG-Transformer 网络很好地解决了 Vison-Transformer 在网络早期进行全局交互时计算量大的问题，将原来的指数计算复杂度优化为线性计算复杂度。

3）在 ImageNet-50 和 ImageNet-100 数据集上进行了大量实验，结果表明 LNG-Transformer 的性能优于 ResNet，Vision Transformer 和 Swin Transformer，并且证明了 LNG-SA 模块的有效性。

2 相关工作

2.1 卷积神经网络

CNN 已经出现几十年，但直到 AlexNet[1] 网络出现，CNN 才开始受到广大研究者的关注，逐渐成为计算机视觉的主流。在此期间，涌现了一大批基于 CNN 的网络，进一步推动了计算机视觉的发展。VGG[18] 使用统一的 3×3 的卷积核来限制每一层的参数数量，通过设计更深的网络深度来增加更大的感受野。GoogLeNet[3] 在层与层之间使用不同大小的卷积核，以提取不同尺度的特征。ResNet[17] 使用恒等映射，增加神经网络架构的残差结构，解决了网络的梯度消失、梯度爆炸和网络退化问题。ResNeXt[19] 借鉴 ResNet 思想和分组卷积，在保证计算量和参数数量与 ResNet 相同的情况下，性能得到提升。DenseNet[5] 使用更激进的恒等映射机制，即互相连接所有的层，每个层都接受其前面所有层作为其额外的输入，来增加网络的学习能力。SE-Net[10] 引入了注意力的思想，即对于每个通道，用一个权重来表示该通道在下一阶段的重要性。MobileNet[6] 采用深度可分离卷积，在保证准确率略微降低的情况下，大幅度减少了网络计算量和参数数量，在移动端和嵌入式领域应用广泛。ShuffleNet[20] 在分组卷积的基础上进行通道重新分组，在减少计算量的同时增加了组间的信息交流。

虽然 CNN 在计算机视觉领域取得了不错的成绩，但如何进一步提升网络性能成为了一个难点。一些学者发现 Transformer 的强大学习能力后，尝试将它应用到视觉任务上，取得了良好的表现，如 Vision Transformer 和 Swin Transformer 等。

2.2 Transformers 在计算机视觉领域的发展与应用

Transformer 最初应用在自然语言处理领域中，由 Transformer 演变出来的 Bert 在 11 种不同自然语言处理测试中达到 SOTA 表现，这是自然语言处理发展史上里程碑式的成就。受 Transformer 和 Bert 的启发，一些学者把 Transformer 应用在计算机视觉领域，例如 Vision Transformer 使用 Transformer 的 Encoder 结构，直接将图像进行窗口划分，处理成一个个 Token 输入到 Encoder 网络，在采用大量训练数据的情况下取得了不错的成绩。但 Self-Attention[9] 的特殊结构导致 Vision Transformer 的参数量巨大，在只有少量训练数据时难以训练。同时，随着输入图像分辨率的增加，Vision Transformer 的计算量呈指数级增长，限制了其工业应用。

随后的 Swin Transformer 提出层次化结构和滑动窗口，很好地解决了 Vision Transformer 参数量大、难以训练等问题。通过把图像划分成一个个大小相同的窗口，只在窗口内进行信息交互，大幅度降低了计算复杂度，与输入图像分辨率呈线性关系。在图像分类、目标检测以及语义分割等任务上，纯视觉的 Transformer 第一次胜过了卷积框架。但窗口化带来了另一个问题，Swin Transformer 不能像 Vision Transformer 一样，在每一层都学到全部信息，只能通过增加网络深度，逐渐合并窗口之间的信息来学习全局信息。为了解决全局信息学习复杂的问题，我们基于 Swin Transformer 提出了一个改进的网络 LNG-Transformer，该网络中的每个模块都包含了局部信息、邻居信息和全局信息的交互。与 Viosn Trans-former 和 Swin Transformer 相比，LNG-Transformer 具有更好的灵活性和更高的效率，在任何一个时期都可以学习到全局信息。在保证近似的计算量和参数规模的情况下，LNG-Transformer 在 ImageNet-50 和 ImageNet-100 图像分类

任务上的 top-1 准确率超过了 ResNet，Vision Transformer 和 Swin Transformer 等网络。

3 LNG-Transformer

3.1 整体结构

图 1 给出了 LNG-Transformer 的整体架构（以最小规模版本 LNG-T 为例，参考 3.4 节），该网络继承了 VGG 网络的层次化结构，主要分为 4 个阶段（Stage），特征图在每一个 Stage 中都会进行局部信息、邻居信息和全局信息的交互，同时对高宽做一次下采样，在特征通道维度上进行翻倍，这样有利于网络提取不同尺度的特征，减少计算量。在不损失全局信息的情况下，LNG-Transformer 降低到了线性计算复杂度，相比 Vsion Transformer 的二次计算复杂度，计算量明显减少。

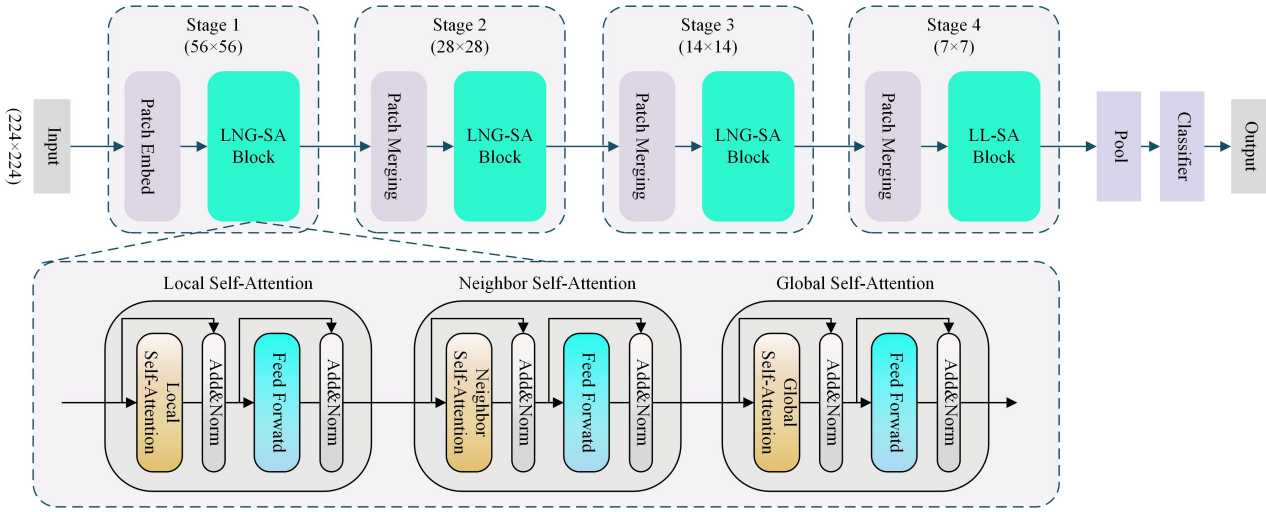


图 1 LNG-Transformer 整体结构

Fig. 1 LNG-Transformer architecture

网络输入图像的高、宽、通道（$H\times W\times C$）为 $224\times224\times3$，Stage 1 中，首先使用 4×4 的窗口，将输入图像划分成一个个特征向量（$4*4*3=48$），再通过线性嵌入层（Patch Embed）将其投影到维度为 96 的向量，得到形状为 $56\times56\times96$ 的特征图，最后通过 LNG-SA 模块得到 Stage 1 的输出，其中 LNG-SA 模块的输入和输出特征图形状不发生变化。Stage 2 中，首先通过线性融合层（Patch Merging）使用 2×2 大小的窗口将输入图像划分成一个个（$2*2*96=384$）新的特征向量，同时将特征维度映射到原来的 1/2，即 182，同样通过一个 LNG-SA 模块得到特征图的输出形状为 $28\times28\times182$。Stage 3 的过程与 Stage 2 相同，输出的形状为 $14\times14\times384$。与之前不同的是，Stage 4 中将 LNG-SA 模块更换成 LL-SA 模块，输出变为 $7\times7\times768$，这是因为 Stage 4 的输入特征图高宽大小与窗口大小相同，均为 7×7，因此只需要进行局部信息学习即可。再通过 Pool 层计算出每一个通道的平均值，此时特征图的形状变为 1×768。最后，通过 Classifier 分类层进行分类。

3.2 Local Neighbor Global Self-Attention 模块

Self-Attention 可以自适应地进行全局关注，并通过整个空间特征相似度关系更新特征，但它在空间上是二次计算复杂度，随着空间特征个数的增加，复杂度呈现指数级增长，这也限制了 Vision Transformer 的应用。Swin Transformer 基于窗口的 Self-Attention 从向量个数的角度很好地解决了复杂度高的问题，使得 Self-Attention 对高分辨率图像进行处理变得可行，但窗口化也导致网络在早期无法获取全局信息。

因此，我们必须把特征图进行窗口化，还需要在早期进行全局信息的学习。为了解决这个问题，我们设计了一个新的自注意力模块 LNG-SA，如图 2 所示。

图 2 LNG-SA 模块内部结构

Fig. 2 Internal structure of LNG-SA

LNG-SA 由 Local Self-Attention，Neighbor Self-Attention 和 Global Self-Attention 这 3 个部分组成，这些部分都是基于标准的多头自注意力机制[9]（Multi-headed Self-attention，MSA）它们分别计算局部注意力、邻居注意力和全局注意力，从而更好地捕捉输入数据的局部、邻居和全局特征。在两层 MLP[9] 之间使用 GELU[21] 激活函数，在每个 MSA 模块

和每个 MLP 之前使用 LayerNorm[22]层，在每个模块之后使用残差连接。LNG-SA 模块计算局部信息、邻居信息和全局信息，LL-SA 模块只计算两次局部信息。最后，通过简单层次化堆叠，构建成一个完整网络。

LNG-SA 的第一步是 Local Self-Attention 局部信息的学习。首先，通过 Window Pation1 进行特征图的窗口划分操作，进行局部信息的学习，再通过 Window Reverse1 还原到原始特征图尺寸。

第二步是 Neighbor Self-Attention 邻居信息的学习。进行窗口划分后，采用滑动窗口方式，在划分窗口前将原始特征图通过 Roll Patition[16]操作，将窗口向右下方向滑动窗口大小的二分之一的距离，再进行邻居信息的学习，最后通过 Window Reverse1 和 Roll Reverse[16]还原特征图尺寸。

第三步是 Global Self-Attention 全局信息的学习。如何将全局特征划分到一个窗口内是这一部分的关键问题。特征图必须遵循采样后窗口大小保持不变，因此可以在高维度上每隔 $H/7$ 个特征进行采样，同时在宽维度上每隔 $W/7$ 个特征进行采样。通过 Window Pation1[16]和 Permute Pation 就可以实现以上操作，使每一个窗口内的特征都包含全局信息。最后，通过 Permute Reverse 和 Window Reverse2[16]还原特征图尺寸。

算法 1 描述了 LNG-SA 模块的具体实现过程。window_partition 实现窗口划分，为计算窗口内注意力做准备；window_reverse 实现窗口还原，为下一个窗口划分做准备。

算法 1 LNG-SA 模块实现算法

输入：特征图 B×H×W×C(批量大小×高×宽×通道)

输出：特征图 B×H×W×C(批量大小×高×宽×通道)

window_partition：B×H×W×C → B×N×S×C

window_reverse：B×N×S×C → B×C×H×W

注：N=H/7 * W/7

S=window_H * window_W

target=H×W

```
# Local Attention
/* window_H=7,window_W=7 */
1. x=window_partition(x,window_H,window_W)
2. x=LocalAttention(x)
3. x=window_reverse(x,window_H,window_W,target)

# Neighbor Attention
/* window_H=7,window_W=7 */
4. x=torch.roll(x,shifts,dims=(1,2)) /* shift_size=-3 */
5. x=self.window_partition(x,window_H,window_W)
6. x=self.NeighborAttention(x,type='N')
7. x=self.window_reverse(x,window_H,window_W,target)
8. x=torch.roll(x,shifts,dims=(1,2)) /* shift_size=3 */

# Global Attention
/* window_H=int(H/7),window_W=int(W/7) */
9. x=window_partition(x,window_H,window_W)
10. x=x.permute(0,2,1,3)
11. x=GlobalAttention(x)
12. x=x.permute(0,2,1,3)
13. x=window_reverse(x,window_H,window_W,target)
```

在 Local Attention 中，窗口划分后，即可以进行局部信息的学习，最后将特征图还原成初始特征图尺寸。

在 Neighbor Attention 中，与 Local Attention 不同的是，首先需要将窗口向右下滑动，滑动的步数为窗口的一半，然后再进行窗口划分以及邻居信息的学习，最后还原成初始特征图尺寸。

在 Global Attention 中，首先将特征图按照 $(H/7)\times(W/7)$ 窗口划分，再进行窗口个数和窗口大小维度的交换，实现在高的维度上每隔 $H/7$ 次做一次采样，在宽的维度上每隔 $W/7$ 做一次采样，并且保证了当前注意力窗口与之前注意力窗口的尺寸不变，即 7×7。然后进行全局信息的学习，最后将特征图还原到初始特征图尺寸。

表 1 LNG-Transformer 结构变量

Table 1 Structural variables of LNG-Transformer

Stage	Downsp rate	LNG-T	LNG-S	LNG-B
Input	1/1	(224×224×3)	(224×224×3)	(224×224×3)
Stage 1	1/4	Patch Embed→(56×56×96) (w=7×7, c=96, h=3)×1	Patch Embed→(56×56×128) (w=7×7, c=128, h=4)×1	Patch Embed→(56×56×128) (w=7×7, c=128, h=4)×1
Stage 2	1/8	Patch Merging→(28×28×192) (w=7×7, c=192, h=6)×1	Patch Merging→(28×28×256) (w=7×7, c=256, h=8)×1	Patch Merging→(28×28×256) (w=7×7, c=256, h=8)×1
Stage 3	1/16	Patch Merging→(14×14×384) (w=7×7, c=384, h=12)×2	Patch Merging→(14×14×512) (w=7×7, c=512, h=16)×2	Patch Merging→(14×14×512) (w=7×7, c=512, h=16)×6
Stage 4	1/32	Patch Merging→(7×7×768) (w=7×7, c=768, h=24)×1	Patch Merging→(7×7×1024) (w=7×7, c=1024, h=32)×1	Patch Merging→(7×7×1024) (w=7×7, c=1024, h=32)×1

注：Downsp rate 为图像高宽下采样倍数；Stage 为网络的不同阶段；w 为窗口大小；c 为通道数；h 为注意力头数。

3.3 结构变量

本文设计了一系列简单的网络结构，整个网络是由 LNG-SA，LL-SA 以及 Classifier 这 3 种模块进行简单拼接组成的，根据规模的不同构建了 3 个网络，分别命名为 LNG-T，LNG-S 和 LNG-B，网络参数和计算量大约是 1，2 和 4 倍。与 Swin Transformer 类似，我们也采用了分阶段的结构。网络主体包含 4 个阶段，每个阶段的注意力窗口大小都为 7×7，Stage 1 对特征图做 4 倍下采样，此后每个阶段都做 2 倍下采样，并且通道数和注意力头数增大为原来的 2 倍。

4 实验

在 ImageNet-50(ImageNet-1k 的前 50 个类别)和 ImageNet-100(ImageNet-1k 的前 100 个类别)图像分类数据集上进行了实验，以 top-1 和 top-5 准确率(top-k 准确率表示预测结果中概率最大的前 k 个结果所包含正确标签的占比)作为评价指标，与一些主流网络进行对比，在此基础上分析 LNG-

Transformer 的设计优点。

4.1 ImageNet-50 和 ImageNet-100 数据集上的图像分类

对于 ImageNet-50 数据集，选取 ImageNet-1k 前 50 类别，训练集共有 64817 张图片，验证集共有 2500 张图片。对于 ImageNet-100 数据集，选取 ImageNet-1k 的前 100 类别，训练集共有 129395 张图片，验证集共有 5000 张图片。具体实验设置如下：在整个训练过程，我们迭代 300 个 epoch，采用 AdamW[23] 优化器，动量参数设置为 0.9，使用余弦衰减学习率调度器和 20 个周期的线性预热。批量大小参照表 2 设置，初始学习率为 5×10^{-4}，最小学习率为 5×10^{-6}，使用 5×10^{-2} 的权重衰减，学习率衰减周期间隔 30 epoch。反向求解梯度过程中，若梯度值超过 5，则做梯度剪切。把网络每一层都初始化为高斯分布，其中均值为 0、方差为 0.02。输入图像的分辨率为 224×224，并且使用旋转、剪切、mixup[24]、色彩抖动等数据增强方法，使用 4 张 GeForce RTX 3090 并行训练。

实验结果如表 2 所列。在 ImageNet-50 数据集上，与以往最先进的基于 Transformer 的体系结构相比，LNG-Transformer 超越了 Swin Transformer 和 Vison Transformer。在 top-1 评价指标上，LNG-T，LNG-S 和 LNG-B 比 Swin-T，Swin-S 和 Swin-B 分别高出 1.04%，1.03%和 0.32%。LNG-T 与 Vit-B 相比高出 8.56%，与卷积神经网络 ResNet-50 相比高出 1.48%。LNG-S 与 ResNet-101 相比，高出 1.62%。

在 ImageNet-100 数据集上，与以往最先进的基于 Transformer 的体系结构相比，LNG-Transformer 超越了 Swin Transformer 和 Vison Transformer。在 top-1 评价指标上，LNG-T，LNG-S 和 LNG-B 比 Swin-T，Swin-S 和 Swin-B 分别高出 0.4%，0.08%和 0.08%。LNG-T 与 Vit-B 相比高出 6.56%，与卷积神经网络 ResNet-50 相比高出 0.8%。LNG-S 与 ResNet-101 相比高出 0.19%。

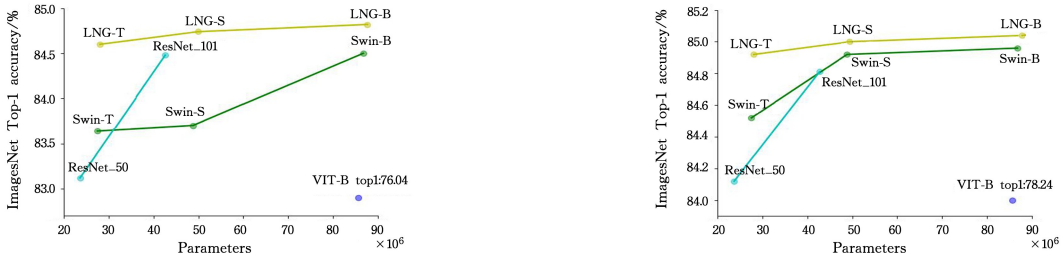
在 top-5 评价指标上，LNG-Transformer 在两个数据集上的表现接近当前主流网络，其中 LNG-B 在 ImageNet-50 数据集上比 Swin-T 提升了 0.16%，LNG-T 在 ImageNet-100 数据集上的性能表现与 Swin-T 相当。

图 3 给出了 Vison Transformer，Swin Transformer，ResNet 在 ImageNet50 数据集和 ImageNet100 数据集上 top-1 准确率与模型参数的曲线图。可以看出，在保证参数量接近的情况下，LNG-Transformer 在 top-1 准确率评价指标上高于现在的主流方法。

表 2 不同网络的参数规模、计算量、top-1 准确率以及 top-5 准确率实验结果对比

Table 2 Experimental results of parameter scale, computation amount, top-1 accuracy and top-5 accuracy of different networks

Dataset	Model	Batch Size	Params	FLOPs	top-1 acc./%	top-5 acc./%
ImageNet-50	LNG-T	210	28.065×10^6	5.044×10^9	**84.60**	96.80
	Swin-T[16]	256	27.510×10^6	4.350×10^9	83.64	**97.00**
	ResNet-50[4]	480	23.610×10^6	4.110×10^9	83.12	96.56
	VIT-B[15]	160	85.647×10^6	16.848×10^9	76.04	92.36
	LNG-S	150	49.907×10^6	8.960×10^9	**84.73**	97.16
	Swin-S[16]	160	48.785×10^6	8.512×10^9	83.70	96.76
	ResNet-101[4]	300	42.603×10^6	7.832×10^9	84.48	**97.40**
	LNG-B	100	87.660×10^6	16.359×10^9	**84.82**	**96.74**
	Swin-B[16]	110	86.725×10^6	15.126×10^9	84.50	96.58
ImageNet-100	LNG-T	210	28.104×10^6	5.044×10^9	**84.92**	**96.82**
	Swin-T	256	27.548×10^6	4.350×10^9	84.52	**96.82**
	ResNet-50	480	23.713×10^6	4.110×10^9	84.12	96.54
	VIT-B	160	85.685×10^6	16.848×10^9	78.26	93.84
	LNG-S	140	49.358×10^6	10.810×10^9	**85.00**	96.72
	Swin-S	160	48.823×10^6	8.512×10^9	84.92	**96.85**
	ResNet-101	300	42.705×10^6	7.832×10^9	84.81	97.08
	LNG-B	100	87.711×10^6	16.359×10^9	**85.04**	96.74
	Swin-B	110	86.763×10^6	15.126×10^9	84.96	**96.88**



(a) ImageNet-50 数据集　　(b) ImageNet-100 数据集

图 3 LNG-Trainsformer 与经典的视觉网络的性能比较

Fig. 3 Performance comparison of LNG-Trainsformer and other classical vision networks

4.2 消融实验

为了证明 LNG-SA 模块的有效性，本文进行了两个消融实验。1）去掉 LNG-SA 模块内部的不同注意力模块进行实验结果的对比，证明 LNG-SA 模块整体的有效性。2）将 LNG-SA 模块迁移到其他网络中进行实验结果的对比，证明 LNG-SA 模块在其他网络中的有效性。

为了证明 LNG-SA 模块整体的有效性，我们重新设计网络，分别去掉 Local Self-Attention 模块、Neighbor Self-At-

tention 模块和 Global Self-Attention 模块进行实验。在 ImageNet-50 和 ImageNet-100 数据集上，使用相同的训练配置和参数规模训练 300 个 epoch，并记录 top-1 和 top-5 的准确率。

实验结果如表 3 所列。在 ImageNet-50 数据集上，本文设计的 LNG-SA 模块相比 no-Local，no-Neighbor 和 no-Global 在 top-1 指标上分别高 2.26%，1.44%，0.98%。在 ImageNet-100 数据集上，本文设计的 LNG-SA 模块相比 no-Local，no-Neighbor 和 no-Global 在 top-1 指标上分别高 2.2%，1.05%，0.39%。结果表明，本文设计的 LNG-SA 模块是有效的，可以在任意时期进行局部注意力、邻居注意力和全局注意力的学习。

表 3 LNG-SA 模块的有效性验证

Table 3 Validity verification of LNG-SA

Dataset	Model	Config	Batch Size	Params	FLOPs	top-1 acc./%	top-5 acc./%
ImageNet-50	LNG-T	[1,1,2,1]	210	28.065×10^6	5.044×10^9	**84.60**	96.80
	no-Local	[1,1,3,1]	256	27.510×10^6	4.350×10^9	82.34	94.78
	no-Neighbor	[1,1,3,1]	256	27.510×10^6	4.350×10^9	83.16	96.56
	no-Global	[1,1,3,1]	160	27.510×10^6	4.350×10^9	83.62	**96.96**
ImageNet-100	LNG-T	[1,1,3,1]	210	28.104×10^6	5.044×10^9	**84.92**	**96.82**
	no-Local	[1,1,3,1]	256	27.548×10^6	4.350×10^9	82.72	94.74
	no-Neighbor	[1,1,3,1]	256	27.548×10^6	4.350×10^9	83.87	96.19
	no-Global	[1,1,3,1]	256	27.548×10^6	4.350×10^9	84.53	96.80

注：no-Local 表示去掉局部注意力模块；no-Neighbor 表示去掉邻居注意力模块；no-Global 表示去掉全局注意力模块。实验根据不同网络重新设置 Config，以保证网络参数数量接近。

为了证明 LNG-SA 模块在其他网络中的有效性，将 LNG-SA 模块加入 Swin-Transformer 和 ResNet 网络中进行实验结果的对比，新网络分别命名为 Swin-X-LNG 和 ResNet-X-LNG(X 表示不同版本)。Swin-X-LNG 网络是在 Swin-Transformer 的 Stage1 结束时重复加入两个 LNG-SA 模块。ResNet-X-LNG 网络是在 ResNet 的 Maxpool 层后重复加入 3 个 LNG-SA 模块。在 ImageNet-50 和 ImageNet-100 数据集上，使用相同的训练配置和参数规模训练 300 个 epoch，并记录 top-1 和 top-5 的准确率。

实验结果如表 4 所列。参考 top-1 指标，在 ImageNet-50 数据集上，Swin-X-LNG 相比 Swin-T，Swin-S 和 Swin-B 分别高 0.55 %，0.69 %，0.17%。ResNet-X-LNG 比 ResNet-50 和 ResNet-101 分别高 0.20%和 0.12%。在 ImageNet-100 数据集上，Swin-X-LNG 相比 Swin-T，Swin-S 和 Swin-B 分别高 0.23%，0.06%，0.03%。ResNet-X-LNG 相比 ResNet-50 和 ResNet-101 分别高出 0.15%和 0.13%。结果表明，我们设计的 LNG-SA 模块在其他网络上是有效的，可以帮助网络在早期进行全局信息的交互。

表 4 LNG-SA 模块在其他网络中的有效性验证

Table 4 Validity verification of LNG-SA in other networks

Dataset	Model	Batch Size	Params	FLOPs	top-1 acc./%	top-5 acc./%
ImageNet-50	Swin-T	256	27.510×10^6	4.350×10^9	83.64	**97.00**
	Swin-T-LNG	230	28.869×10^6	5.418×10^9	**84.19**	96.68
	Swin-S	160	48.785×10^6	8.512×10^9	83.70	**96.76**
	Swin-S-LNG	130	51.497×10^6	10.638×10^9	**84.39**	96.68
	Swin-B	110	86.725×10^6	15.126×10^9	84.50	96.58
	Swin-B-LNG	100	89.100×10^6	17.026×10^9	**84.67**	**97.32**
	ResNet-50	480	23.610×10^6	4.110×10^9	83.12	96.56
	ResNet-50-LNG	450	24.060×10^6	5.566×10^9	**83.32**	**96.67**
	ResNet-101[4]	300	42.603×10^6	7.832×10^9	84.48	**97.40**
	ResNet-101-LNG	280	43.052×10^6	9.299×10^9	**84.60**	**97.40**
ImageNet-100	Swin-T	256	27.548×10^6	4.350×10^9	84.52	96.82
	Swin-T-LNG	230	30.242×10^6	6.465×10^9	**84.75**	**97.11**
	Swin-S	160	48.823×10^6	8.512×10^9	84.92	96.85
	Swin-S-LNG	130	51.536×10^6	10.638×10^9	**84.98**	**97.15**
	Swin-B	110	86.763×10^6	15.126×10^9	84.96	96.88
	Swin-B-LNG	100	91.521×10^6	18.883×10^9	**84.99**	**97.23**
	ResNet-50	480	23.713×10^6	4.110×10^9	84.12	**96.54**
	ResNet-50-LNG	450	24.163×10^6	5.567×10^9	**84.27**	97.31
	ResNet-101	300	42.705×10^6	7.832×10^9	84.81	**97.08**
	ResNet-101-LNG	280	43.155×10^6	9.299×10^9	**84.94**	97.02

结束语 本文提出了一种有效的、可扩展的注意力模块 LNG-SA，该模块在任意时期都能进行局部信息、邻居信息和全局信息的交互。通过重复级联 LNG-SA 模块，设计了一个全新的网络，称为 LNG-Transformer。该网络整体采用层次化结构，具有优秀的灵活性，其计算复杂度与图像分辨率呈线性关系。LNG-SA 模块的特性使得 LNG-Transformer 即使在早期的高分辨率阶段也可以进行局部信息、邻居信息和全局信息的交互，从而带来更高的效率和更强的学习能力。在 ImageNet-50 和 ImageNet-100 数据集上，所提方法超过了现有的主流方法。并且将 LNG-SA 模块加入 Swin-Transformer 和 ResNet 网络中，都能提升图像分类的 top-1 的准确率，也证明了 LNG-SA 模型的泛化能力。

参 考 文 献

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

[2] DING X，ZHANG X，HAN J，et al. Scaling up your kernels to 31×31：Revisiting large kernel design in cnns[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022：11963-11975.

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

[4] 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.

[5] HUANG 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.

[6] ANDREW G H，MENGLONG Z，BO C，et al. MobileNets：Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. arXiv：1704. 04861，2017.

[7] FISHER Y，VLADLEN K. Multi-Scale Context Aggregation by Dilated Convolutions[J]. arXiv：1511. 07122，2015.

[8] DING X，ZHANG X，MA N，et al. Repvgg：Making vgg-style convnets great again[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：13733-13742.

[9] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing SystemsDecember. 2017：6000-6010.

[10] 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.

[11] HAN X L，CHEN J C，ZHOU W S. New MobileNet image classification algorithm based on 3D attention[J]. Journal of Chongqing University of Post and Telecommunications(Natural Science Edition)，2023，35(3)：513-519.

[12] ZHAO H W，ZHANG J R，ZHU J P，et al. Image classification framework based on contrastive self-supervised learning[J]. Journal of Jilin University(Engineering and Technology Edition)，2022，52(8)：1850-1856.

[13] 张辉宜，夏媛龙，周克武，等. 一种融合标签间强相关性的多标签图像分类方法[J]. 重庆工商大学学报(自然科学版)，2023，40(5)：8-15.

[14] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv：1810. 04805，2019.

[15] ALEXEY D，LUCAS B，ALEXANDER K，et al. An Image is Worth 16×16 Words：Transformers for Image Recognition at Scale[C]// International Conference on Learning Representations. 2021.

[16] LIU Z，LIN Y，CAO Y，et al. Swin transformer：Hierarchical vision transformer using neighbored windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：10012-10022.

[17] 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.

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

[19] 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.

[20] ZHANG X，ZHOU X，LIN M，et al. Shufflenet：An extremely efficient convolutional neural network for mobile devices[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：6848-6856.

[21] DAN H，KEVIN G. Gaussian Error Linear Units(GELUs)[J]. arXiv：1606. 08415，2016.

[22] LEI J B，RYAN K，GEOFFREY E H，et al. Layer Normalization[J]. arXiv：1607. 06450，2016.

[23] ILYA L，FRANK H. Decoupled Weight Decay Regularization[J]. arXiv：1711. 05101，2019.

[24] HONGYI Z，MOUSTAPHA C，YANN N D，et al. mixup：Beyond Empirical Risk Minimization[J]. arXiv：1710. 09412，23018.

WANG Wenjie，born in 1996，postgraduate. His main research interests include image processing andobject detection.

YANG Yan，born in 1964，professor，Ph.D supervisor，is a member of CCF (No. 06877D). Her main research interests include multi-view learning，cluster analysis and so on.

（责任编辑：何杨）