将图像压缩成用于视觉转换器的图像块

赵鑫峰，孙耀儒

同济大学

Abstract

视觉Transformer (ViT) 在计算机视觉领域取得了显著进展。然而，随着模型深度的增加和输入图像分辨率的提高，训练和运行ViT模型的计算成本急剧增加。本文提出了一种基于CNN和视觉Transformer的混合模型，命名为CI2P-ViT。该模型包含一个名为CI2P的模块，该模块利用CompressAI编码器压缩图像，然后通过一系列卷积生成一系列patch。CI2P可以替代ViT模型中的Patch Embedding组件，从而无缝集成到现有的ViT模型中。与ViT-B/16相比，CI2P-ViT将输入到自注意力层的patch数量减少到原来的四分之一。这种设计不仅显著降低了ViT模型的计算成本，而且通过引入CNN的归纳偏置特性有效地提高了模型的精度。ViT模型的精度显著提高。在Animals-10数据集上从头开始训练时，CI2P-ViT达到了92.37%的准确率，比ViT-B/16基线提高了3.3%。此外，该模型的计算操作（以每秒浮点运算次数（FLOPs）衡量）减少了63.35%，并且在相同的硬件配置下训练速度提高了2倍。

1 Introduction

在计算机视觉领域，如[1, 2, 3, 4]中所述，卷积神经网络（CNN）通过局部感受野和权值共享有效地捕获局部图像特征，在各种计算机视觉任务中展现出卓越的性能。

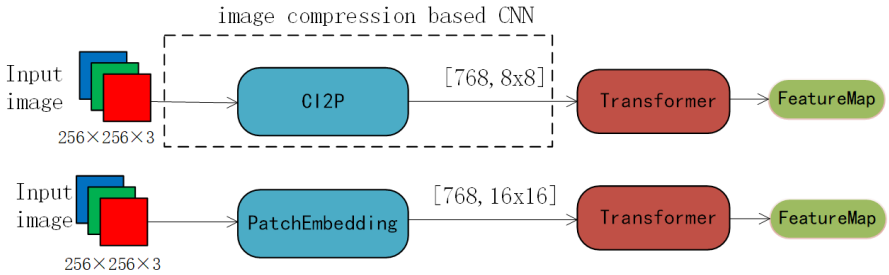

Fig. 1CI2P模块的作用类似于标准ViT架构中的Patch Embedding组件，但它生成的Patch大小只有原始Patch的四分之一，从而最大限度地减少了视觉保真度的损失。因此，这项创新使ViT-B/16模型的FLOPs降低了63.35%（对于256x256分辨率的图像）。

视觉任务。然而，由于CNNs感受野有限的特点，它们在捕捉图像中的全局特征信息和建立长程依赖关系方面存在局限性。

Transformer架构，正如[5, 6]中所述，彻底改变了自然语言处理领域，其核心机制是自注意力机制，该机制擅长捕获序列数据中的全局依赖关系。在[7]中提出的ViT成功地将这种自注意力机制集成到图像处理中，展现出其独特的优势和巨大的潜力。具体而言，在需要全面理解图像结构和复杂关系动态的任务中，ViT表现出卓越的性能。然而，由于缺乏局部相关性和平移不变性等归纳偏置，ViT在与CNN相比处于劣势，导致训练数据需求增加。此外，ViT缺乏下采样机制，导致计算成本更高，以FLOPs衡量。针对ViT在图像任务中计算效率低和资源需求大的问题，研究人员设计了多种ViT变体，旨在融合ViT和CNN的优势，从而提高模型的性能和效率。

ViT将图像分割成16x16像素的网格，并应用自注意力机制来处理图像数据。自注意力层（MSA）的运算复杂度与图像尺寸和模型容量成正比，如公式[8]所示：{v*}

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C \tag{1}$$

这里，C表示每个图像块的维度，hw是将其分割成块后的结果大小。为了减轻ViT的计算需求，可以调整模型的容量C或图像块的空间范围hw。减小C可能会损害模型捕获复杂特征的能力，因此典型的策略是缩小尺寸

图像大小hw。Swin Transformer [8]就是这种优化的一个例子，它采用分层结构和移位窗口机制，通过增量调整每一层的hw和C来提高性能。CNN-Transformer混合模型利用CNN的下采样技术来减小空间维度hw，同时放大通道数，从而降低自注意力机制的计算开销。例如，使用基于CNN的Backbone（例如ResNet [9]）生成的特征图，并将它们展平作为ViT的输入，可以显著减少计算成本。然而，这种方法存在权衡：它虽然最小化了图像数据的维度，但也可能导致重要视觉信息的丢失。CNN-Backbone通常在ImageNet分类任务上进行预训练，以培养对图像内容的深刻理解，从而产生富含语义信息的特征图。然而，这些特征图可能会忽略原始图像中的关键视觉细节，当试图从这些特征图重建原始图像时，这种缺点就会显现出来。对于Vision Transformer，这种被忽略的信息可能是至关重要的，因为它可能在处理原始图像的过程中捕获独特的语义特征。在像素密集型任务（例如小目标检测、语义分割和姿态估计）中，细粒度细节的保存至关重要，这种视觉信息的丢失尤其有害。

这项研究介绍了一种突破性的混合模型CI2P-ViT，它结合了CNN和Transformer的优势。CI2P模块采用源自CNN的图像压缩技术来降低输入到ViT的维度，有效地将自注意力层的计算FLOPs减少了四分之一。如图1所示，这种减少是在不影响模型捕捉图像中所有视觉细节的能力的情况下实现的。

CI2P模块的降维组件是独立于特定视觉任务进行训练的，它使用编码器-解码器框架。这种方法确保图像首先被编码器压缩成低维表示，然后由解码器精确重建，尽可能保留原始视觉信息。

此外，CI2P模块将CNN的归纳偏差整合到ViT框架中，从而增强模型检测和处理图像中局部特征的能力。CI2P-ViT的一个关键优势在于它保留了原始ViT架构，与Swin Transformer等需要修改ViT内部结构的其他模型相比，使其更易于适应多模态研究和扩展。

这种设计理念不仅简化了模型的计算效率，也拓宽了其适用性，为未来多模态分析及其他领域的创新铺平了道路。

本文的主要贡献包括以下内容：

- 我们提出了一种基于CNN的CI2P模块，它可以作为ViT的Patch Embedding阶段的即插即用组件。这种方法

在保持ViT模型结构的同时，融合了CNN的归纳偏置。

- 我们引入了一种开创性的方法，将基于CNN的图像压缩技术CompressAI与ViT框架融合。这种协同作用降低了图像数据的维度，从而显著减少了模型的FLOPs。
- 我们提出了一种名为CI2P-ViT的创新混合模型。在此基础上，设计了一种具有双尺度注意力机制的变体CI2P-ViT^{ds}。我们的模型的有效性通过图像分类数据集上的实验结果得到了证实。

我们的模型在不改变底层ViT结构的情况下，提高了性能和精度。在Animals-10数据集上从零开始训练后，我们的模型超越了基线ViT模型，实现了92.37%的准确率，相当于提高了3.3%。该模型还将FLOPs降低了63.35%，从而显著减少了ViT模型执行所需的内存占用。这一进步对于资源受限的研究人员尤其有利，因为它能够进行更广泛的优化，并拓宽了ViT模型进一步发展的潜力。

2 Related Works

在深度学习领域，文献[1]中介绍的卷积神经网络（CNN）已被确立为各种计算机视觉任务的基础架构。CNN通过其卷积层有效地捕获局部图像特征，在图像分类[2]、目标检测[3]和语义分割[4]方面取得了显著成就。尽管取得了这些成功，但CNN在捕获远程依赖关系和整合图像中的全局上下文信息方面仍然面临挑战。

文献[7]中详细介绍的ViT，通过采用NLP中Transformer模型[5]的机制并将其应用于计算机视觉领域，代表了一种新颖的架构转变。ViT通过将图像分割成patch并将这些patch作为Transformer的序列输入来处理图像，并在各种视觉任务中展现出显著的性能。本综述将简要考察与CNN和视觉Transformer相关的模型，重点介绍它们的演变、应用以及在计算机视觉研究领域中的相互作用。

2.1 Vision Transformer

如[5]中所述，Transformer模型在NLP领域取得了显著成功。这一成功激励研究人员探索Transformer概念在视觉任务中的应用。Dosovitskiy等人[7]提出了视觉Transformer (ViT)，这是一种用于图像识别的模型。ViT的工作原理是将图像分割成一系列patch，然后

通过Transformer架构将图像处理为序列数据，从而在图像分类任务中取得了优越的性能。尽管取得了这些进展，但与CNN相比，ViT的训练需要更大的数据集和更强大的计算资源。DeiT（数据高效图像Transformer）[10]通过整合注意力训练和知识蒸馏技术解决了这一挑战，证明了仅使用ImageNet数据集即可有效训练ViT。DeiT利用预训练的CNN（例如ResNet）采用师生配置，其中CNN提供指导以增强Transformer的自注意力机制，从而提高整体模型性能。Swin Transformer是ViT的改进版本，它引入了一种新颖的分层结构和移位窗口机制，使其区别于其前身[8]。该模型通过实现局部窗口自注意力机制优化了计算效率，该机制将自注意力计算限制在局部图像窗口内，从而显著减少了计算需求。Swin Transformer在计算机视觉领域取得了显著成果，在目标检测、语义分割、图像生成和视频动作识别等多种任务中取得了优异的成果，展示了其在处理复杂视觉挑战方面的多功能性和鲁棒性。

2.2 CNN-Transformer Hybrid Model

CNN-Transformer混合模型是一种创新的神经网络架构，它融合了CNN在局部特征提取方面的优势以及Transformer模型在全局上下文理解方面的全面能力。这些模型旨在利用这两种架构的优势，从而显著提升视觉任务的性能。

在上下文视觉转换器（CvT）[11]中，作者提出了一种独特的结构化方法，在每个注意力机制之前加入一系列卷积层。这种设计有效地降低了各层特征图的空间分辨率，同时扩展了它们的特征维度。作者报告说，与ViT和DeiT相比，CvT在ImageNet-1k数据集上取得了更好的性能，同时参数数量更少，计算需求也更低。

相反，交叉编码图像变换器 (CeiT) [12] 采用了一种替代策略，利用卷积层进行初始图像下采样，然后再通过 ViT 进行处理。这种方法利用了 CNN 在低级特征提取方面的优势，并通过减少经受自注意力机制的图像块序列长度来减少计算开销。

2.3 End to End Image Comparison

CompressAI，在[13]中介绍，是一个基于PyTorch的库和评估框架，精心设计以促进端到端图像压缩研究。该平台包含多个最先进的端到端图像

利用CNN的压缩模型[14, 15, 16]。这些模型已经证明了与已建立的JPEG和PNG算法竞争的压缩能力，从而突出了深度学习在推进图像压缩技术方面的广阔前景。

2.4 Comparison

ViT和DeiT将原始尺寸的图像作为自注意力层的输入，这导致随着模型复杂度和图像分辨率的提高，计算需求显著增加。虽然Swin Transformer通过其开创性的局部窗口机制有效地减少了FLOPs数量，但其对ViT内部结构的修改可能会阻碍其在多模态研究和可扩展性方面的适应性。其他模型，例如CvT、CeiT和其他基于CNN嵌入的架构，采用了利用从CNN提取的特征图作为自注意力层输入的技术，从而降低了计算成本。然而，这种方法固有的下采样可能会导致精细的视觉信息丢失，而这对于需要高图像保真度的任务（例如语义分割和姿态估计）至关重要。在这些像素密集型应用中，图像的细微之处尤其关键。

CNN-Embedding范式中的各种方法通常结合从预训练的CNN架构（如ResNet-50）中提取的特征图，这些架构通常最初是在ImageNet数据集上训练的。这些模型推广到ImageNet之外的数据集时的有效性可能不一致，因此使得ViT模型的性能在很大程度上取决于CNN-Embedding训练过程的熟练程度。此外，CNN-Embedding部分通常需要在ViT训练阶段进行微调。此过程增加了计算成本，并使优化改进的归因复杂化，模糊了改进是源于CNN-Embedding部分还是ViT架构本身。这种可变性进一步使ViT新结构的追求和消融研究的执行复杂化。

在本研究中，我们提出了CI2P-ViT模型，该模型集成了源自CNN的端到端压缩技术，对输入图像进行下采样，有效降低图像维度，同时努力保留重要的视觉信息。这种创新方法使ViT能够更细致地理解图像中的复杂细节。

与通常针对特定视觉任务而定制的传统CNN嵌入技术不同，CI2P的图像压缩组件是自主训练的，从而增强了其通用性。CI2P-ViT模型训练方案的一个显著特点是冻结图像压缩模型的参数。这一策略性决策最大限度地减少了训练相关的计算成本。

我们采用图像压缩技术，将ViT自注意力层处理的图像块数量减少了四倍，从而显著降低了模型的计算负载（以FLOPs衡量）。CI2P模块的集成将CNN归纳偏差引入模型，提高了模型精度并加快了训练速度。与Swin Transformer和其他CVT变体等替代模型相比，CI2P-ViT模型通过保留ViT原始的内部架构而与众不同，避免了任何修改。这种策略促进了Transformer框架在多模态研究领域的更广泛应用和扩展。

3 Method

我们提出了一种名为CI2P-ViT的创新模型，其结构布局如图2所示。CI2P模块包含两个关键组件：CI2P-Encoder和CI2P-PatchReshape。CI2P-Encoder充当图像压缩单元，集成了一个源自CompressAI[13]的有损压缩模型。它利用CNN的能力生成紧凑的潜在表示，该表示与原始图像的视觉质量非常相似。值得注意的是，该组件的训练与ViT模型的训练是解耦的，在整个ViT训练过程中，CI2P-Encoder的参数保持冻结。这种设计策略保证了模型能力的改进本质上是由于ViT的内在优化。CI2P-PatchReshape模块充当维度调整组件，负责将压缩后的图像输出与ViT模型预期的输入维度对齐。

3.1 Compress Image to patches

CI2P框架中的CI2P-Encoder组件负责图像压缩，它使用CompressAI库，该库提供了一套基于CNN的端到端图像压缩算法。在CompressAI框架中，损失函数定义如下：

$$y_{[\frac{c*d*d}{s},\frac{h}{d},\frac{w}{d}]} = encoder(x_{[c,h,w]}) \tag{2}$$

$$\hat{y}_{[\frac{c*d*d}{s},\frac{h}{d},\frac{w}{d}]}, y^* = quantize(y_{[\frac{c*d*d}{s},\frac{h}{d},\frac{w}{d}]}) \tag{3}$$

$$\hat{x}_{[c,h,w]} = decoder(\hat{y}_{[\frac{c*d*d}{s},\frac{h}{d},\frac{w}{d}]}) \tag{4}$$

$$\mathcal{D}_{MSE} = (x_{[c,h,w]} - \hat{x}_{[c,h,w]})^2 \tag{5}$$

$$\mathcal{R}_{bpp} = entropy(y^*) \tag{6}$$

$$\mathcal{L} = \lambda \mathcal{D}_{MSE} + \mathcal{R}_{bpp} \tag{7}$$

在CompressAI提供的模型中，CI2P选择bmshj2018分解(quality=5)模型，如[15]所述，仅使用

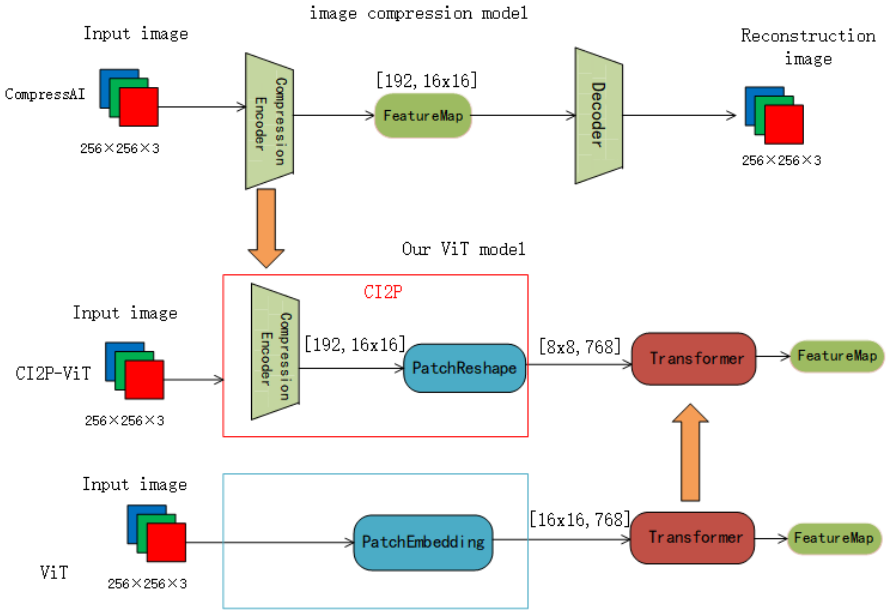

Fig. 2CI2P充当ViT的Patch Embedding模块。编码器部分来自图像压缩模型。

其编码器组件。这种策略性选择是为了在图像压缩效率和视觉保真度之间取得最佳平衡。

损失函数$\mathcal{L}$(7)由两个部分组成：$\mathcal{D}_{MSE}$，它计算原始图像和重建图像之间的均方误差；以及$\mathcal{R}_{bpp}$，它评估编码比特率的效率。编码器，如公式(2)所示，产生原始图像$x_{[c,h,w]}$的降采样表示$y_{[\frac{c*d*d}{s},\frac{h}{d},\frac{w}{d}]}$，其中$c$是颜色通道数，$h$和$w$分别是图像的高度和宽度，$d$是设置为32的降采样维度。缩减因子$s$（本文中为4）决定图像尺寸减小的程度。解码器能够逆转这个过程，从压缩表示$y_{[\frac{c*d*d}{s},\frac{h}{d},\frac{w}{d}]}$重建原始图像，从而确保原始图像的所有视觉信息都得到保留。

CI2P模块的CI2P-PatchReshape组件处理必要的维度重塑，以使压缩图像数据与ViT模型的输入要求对齐。

$$x_{[N,D]} = Flatten(PatchReshape(y_{[\frac{c*d*d}{s},\frac{h}{d},\frac{w}{d}]})) \tag{8}$$

bmshj2018 因式分解模型的编码器产生一个维度为 [192, 16x16] 的输出张量$y_{[\frac{c*d*d}{s},\frac{h}{d},\frac{w}{d}]}$，这与

ViT-B/16模型所需的输入维度，具体为768。为了解决这种不匹配，采用了CI2P-PatchReshape组件来相应地调整维度。卷积后，空间维度减半，通道维度增加四倍，产生[768, 8x8]的输出。然后将此重塑后的输出展平，以匹配ViT模型的输入要求。在CompressAI库中提供的各种模型中，bmshj2018分解模型因其在压缩性能和计算效率之间具有优越的平衡性而被选中。预计未来的研究将产生能够提供更高压缩效率和更低视觉损失的深度学习编码器，从而进一步增强CI2P等模型在未来应用中的能力。

3.2 CI2P-ViT

与传统的ViT架构相比，CI2P模块的集成只需要简单地替换Patch Embedding组件。CI2P模块的输出，如公式(8)所示，作为自注意力机制的输入，其公式如下：

$$feature\_map = transformer(x_{[N,D]}) \tag{9}$$

这种自注意力机制与ViT-B/16采用的配置相同，其注意力层维度为$D = 768$，包含12个注意力头，MLP隐藏层维度为3072。对于尺寸为256x256的输入图像，原始ViT中馈入自注意力层的序列长度N为256，而在CI2P-ViT模型中，它被减少到64。这种减少有效地最小化了FLOPs。CI2P-ViT的特征提取主干由全局平均池化(GAP)操作完成。随后，应用线性分类器来预测输出。CI2P模块的初始部分，即CI2P-Encoders，已经单独进行了预训练，并在CI2P-ViT训练阶段保持冻结状态。

3.3 CI2P-ViT^{ds} with Dual-Scale Attention Mechanism

与原始CI2P-ViT相比，CI2P-ViT^{ds}结合了双尺度注意力机制。CI2P模块的输出维度设置为[192, 16x16]，允许ViT的前六个注意力层在更大的16x16空间尺度上进行注意力计算。随后，一个反向残差网络单元将维度扩大四倍，同时将宽度和高度减半，导致维度变为[768, 8x8]。接下来的六个注意力层然后在8x8空间尺度上进行注意力计算。架构图如图3所示。PatchReshape和CnnReshape都使用了MobileNet[17]中的反向残差网络单元，如图4所示。

与CI2P-ViT相比，前六个注意力层使用192维，这显著减少了模型的参数。模型参数为

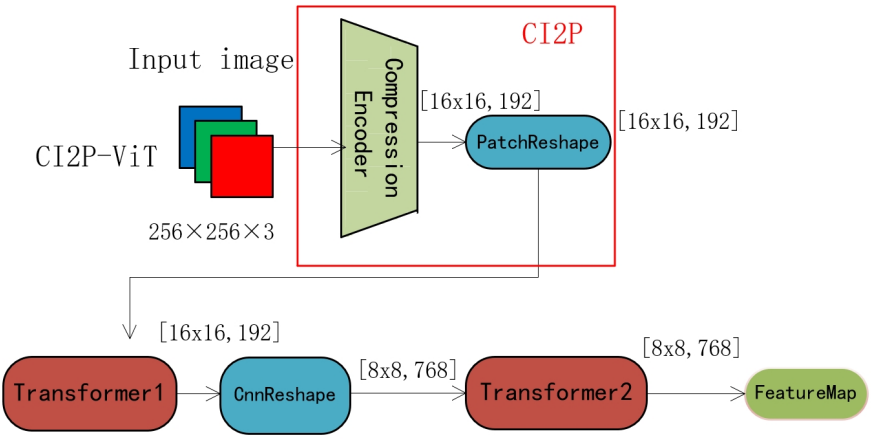

Fig. 3具有双尺度注意力机制的CI2P-ViT^{ds}

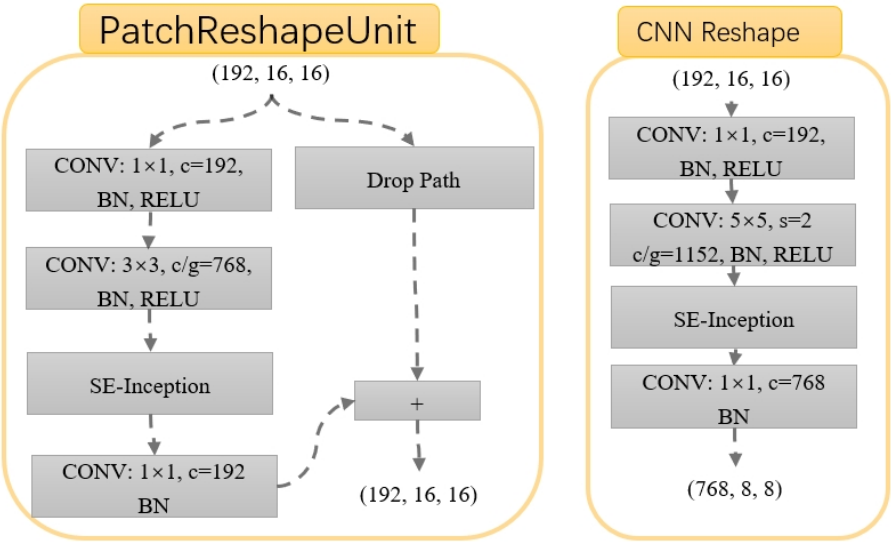

Fig. 4 PatchReshape和CnnReshape。

参数量从8896万减少到4970万，FLOPs从84.77亿减少到64.42亿。同时，Image Net数据集上的精度从72.9%提高到77%，详情见表1。在16x16和8x8两种尺度上进行注意力计算，允许CI2P-ViT^{ds}输出两种尺度的特征图，为诸如目标识别和姿态估计等大量利用多尺度融合技术的任务提供更丰富的语义信息。

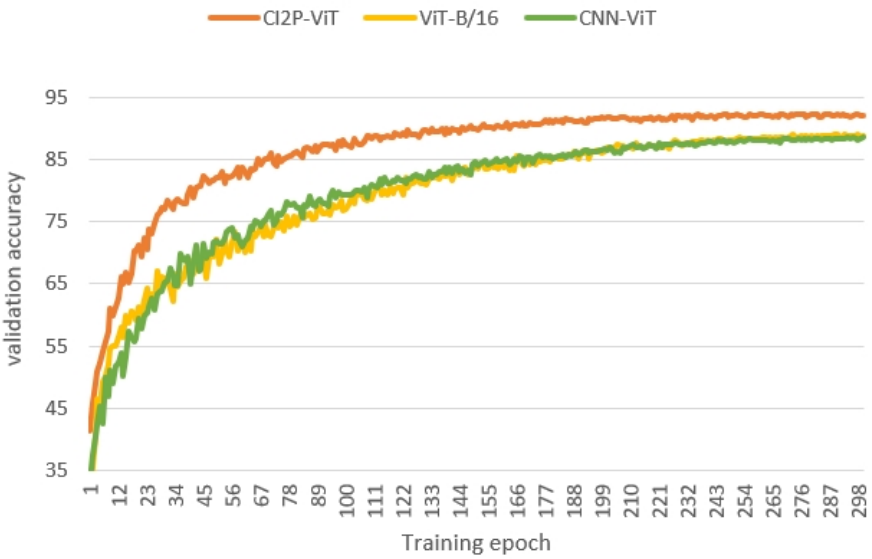

Fig. 5 Animals-10验证准确率。

4 Experiments

4.1 Results on Animals-10 Dataset

在Animals-10[18]数据集上的实验中，我们从零开始训练了CI2P-ViT和ViT-B/16模型。我们采用了mmlab框架。初始学习率设置为1e-04。Adam优化器的参数为$\beta 1 = 0.9$和$\beta 2 = 0.999$。我们使用了概率为0.5的随机翻转变换。在相同的硬件设置和训练参数下，CI2P-ViT在约9小时内完成了300个epoch的训练，而ViT-B/16则需要大约20小时，这表明训练时间显著减少。如图5所示，CI2P-ViT在测试集上达到了92.37%的准确率，优于达到89%准确率的ViT-B/16。

与传统的CNN+ViT模型不同，CI2P模块力求在进行下采样的同时尽可能保留图像中的视觉信息。为了验证CI2P设计的有效性，我们构建了一个对照模型CNN-ViT，它使用与CI2P相同的CNN架构，但不使用来自CompressAI模型的预训练参数。在CNN-ViT中，CNN的所有参数都与ViT模型一起训练。经过300轮训练后，CNN-ViT达到了88.5%的最大准确率，比CI2P-ViT的准确率低3.86%。实验结果表明，仅使用CNN下采样得到的特征图作为ViT的输入不足以保证模型准确率的提高。CI2P模块通过引入CNN的归纳偏置能力并保留

Table 1 ImageNet Top-1 验证准确率

Model	Params(M)	FLOPs(G)	Top-1(%)
DeiT-B [10]	86	17.6	81.8
T2T-ViT-24 [19]	64.1	14.1	82.3
TNT-B [20]	65.6	14.1	82.9
CPVT-B [21]	88	17.6	82.3
Swin-B [8]	88	15.4	83.3
ViT-B/32	86	23.13	73.38
ViT-B/16	86	23.13	77.91
CI2P-ViT	88.96	8.477	72.9
CI2P-ViT^{ds}	49.7	6.442	77

Table 2FLOPs 比较。

Image Size	ViT FLOPs	CI2P-ViT FLOPs	CI2P-ViT^{ds} FLOPs
256^2	23.127 G	8.477(63.35% ↓) G	6.442(72.15% ↓) G
384^2	55.433 G	19.284(65.21% ↓) G	14.492(73.86% ↓) G
512^2	107 G	34.81(67.47% ↓) G	25.762(75.92% ↓) G

图像的视觉信息，可以有效地利用ViT的潜力，从而在视觉识别任务中取得更好的性能。

4.2 Results on the ImageNet Dataset

在ImageNet数据集上经过300轮次的全面训练，CI2P-ViT模型在没有任何预训练数据的情况下，实现了72.9%的测试准确率。与当前主流ViT模型的性能相比，具体对比数据见表1。

尽管CI2P-ViT模型在ImageNet数据集上的准确率低于一些改进的ViT变体，但其性能与最初的ViT-B/32 [7]模型的准确率非常接近。在训练阶段，CI2P-ViT采用了有限的数据增强方法，仅限于随机翻转。它没有在ImageNet-21K等大型数据集上进行预训练，也没有使用更高分辨率的图像进行训练。负责图像块嵌入的CI2P模块使用一个编码器，其参数在整个训练过程中保持不变。虽然该图像编码器在ImageNet上的效力尚未达到峰值，但我们预计随着端到端图像压缩技术的不断发展，CI2P模块的能力将得到显著提升。

4.3 Performance Comparison

ViT-B/16的模型参数为86M，而CI2P-ViT的模型参数为88.96M。由于增加了CI2P模块，模型参数略有增加，但FLOPs显著减少。具体来说，当输入图像大小为256x256像素时，与ViT-B/16相比，CI2P-ViT的FLOPs减少了63.35%。随着

随着模型层数的增加和输入图像尺寸的扩大，CI2P模块带来的FLOPs减少效果越发明显。表2详细说明了这种性能优势。CI2P-ViT^{ds}的模型参数为49.7M，比ViT-B/16减少了42%，FLOPs为6.442G，减少了72.15%，实现了非常好的轻量化效果。

5 Conclusion

本文提出了一种名为CI2P的模块，旨在取代传统ViT架构中的Patch Embedding部分。CI2P模块采用基于CNN的有损压缩技术，在保留几乎所有视觉信息的同时降低图像数据的维度。这种设计显著降低了ViT自注意力层的计算复杂度，而不会牺牲模型的精度。事实上，CI2P模块通过引入CNN的归纳偏置，略微提高了ViT的精度。鉴于计算机视觉应用中图像维度的不断扩大和ViT模型复杂度的不断增加，CI2P技术具有巨大的潜力，表明其在提供更高效和更精确的视觉任务解决方案方面发挥着关键作用。

References

1. Krizhevsky A，Sutskever I，Hinton GE（2017）基于深度卷积神经网络的ImageNet分类。ACM通讯 60(6):84–90

2. LeCun Y，Boser B，Denker JS等（1989）反向传播应用于手写邮政编码识别。神经计算1（4）：541-551

3. 更快的R-CNN (2015) 基于区域建议网络的实时目标检测方法。神经信息处理系统进展 9199(10.5555):2969,239–2969,250

4. Long J，Shelhamer E，Darrell T（2015）用于语义分割的全卷积网络。见：IEEE计算机视觉与模式识别会议论文集，第3431-3440页 5. Vaswani A，Shazeer N，Parmar N，等（2017）注意力是你所需要的一切。神经信息处理系统进展30：5998-6008

6. Devlin J，Chang MW，Lee K，等（2018）Bert：用于语言理解的深度双向Transformer预训练。arXiv预印本arXiv:1810.04805

7. Dosovitskiy A，Beyer L，Kolesnikov A，等 (2020) 一幅图像值16x16个单词：用于大规模图像识别的Transformer。arXiv预印本 arXiv:2010.11929

8. 刘 Z，林 Y，曹 Y，等 (2021) Swin Transformer：基于移位窗口的分层视觉Transformer。见：IEEE/CVF 国际计算机视觉会议论文集，第 10,012−10,022 页

9. 何凯明，张祥雨，任少卿，等 (2016) 深度残差学习用于图像识别。见：IEEE计算机视觉与模式识别会议论文集，第770-778页

10. Touvron H，Cord M，Douze M，等 (2021) 通过注意力机制训练数据高效的图像Transformer & 蒸馏。In: 国际机器学习会议，PMLR，第10347-10357页

11. Hassani A，Walton S，Shah N，等（2021）使用紧凑型转换器摆脱大数据范式。arXiv 预印本 arXiv:2104.05704

12.袁K，郭S，刘Z，等（2021）将卷积设计融入视觉Transformer。见：IEEE/CVF国际计算机视觉会议论文集，第579-588页

13. Begaint J，Racapé F，Feltman S，等（2020）Compressai：一个用于端到端压缩研究的PyTorch库和评估平台。arXiv预印本arXiv:2011.03029

14. Ballé J，Minnen D，Singh S，等 (2018) 具有尺度超先验的变分图像压缩。arXiv 预印本 arXiv:1802.01436

15. Minnen D，Ballé J，Toderici GD（2018）用于学习图像压缩的联合自回归和分层先验。神经信息处理系统进展31

16. 程 Z，孙 H，Takeuchi M，等 (2020) 基于离散高斯混合似然和注意力模块的学习图像压缩。见：IEEE/CVF计算机视觉与模式识别会议论文集，第 7939-7948 页

17. Andrew G，Menglong Z，等 (2017) 用于移动视觉应用的高效卷积神经网络。Mobilenets 10:151

18. Alessio C (2020) Animals-10。网址 https://www.kaggle.com/datasets/alessiocorrado99/animals10/data

19. Yuan L，Chen Y，Wang T，等 (2021) Tokens-to-token vit：从ImageNet上从零开始训练视觉Transformer。见：IEEE/CVF计算机视觉国际会议论文集，第558-567页

20. 韩K，肖A，吴E，等 (2021) Transformer in transformer. Advances in neural information processing systems 34:15,908−15,919

21. Chu X, Tian Z, Zhang B, et al (2021) Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882