卷积神经网络与 ViT 的比较

Koorye

电子科技大学，计算机科学与工程学院

摘　要　在计算机视觉领域，卷积神经网络 (CNN) 和 Vision Transformer 模型 (ViT) 是两种重要的模型。CNN 是一种传统的基于卷积操作的神经网络，在图像分类、目标检测等任务中取得了巨大成功。而 ViT 是一种更为新颖的基于注意力机制的模型，从自然语言处理任务中受到启发，近年来在计算机视觉任务中取得了很好的效果。本文将对这两种模型进行比较，分析它们的优劣势。具体来说，我们将通过比较两者的结构、特征表示、泛化能力等方面，来分析两者的差异。我们将通过基于全局和局部的比较、降维可视化、性能和泛化能力对比等方法，来展示两者的差异。实验结果表明，CNN 存在许多意想不到的优势，其特征区域的形状和模式更接近人类大脑的工作模式，同时 CNN 能更好关注细粒度特征，并具备更强大的泛化能力。希望本文能够为未来的工作带来新的启发。

关键词　卷积神经网络、Vision Transformer。

1　介绍

在计算机视觉领域，卷积神经网络 (CNN)[L] 和 Vision Transformer(ViT)[D] 是两种截然不同的模型，前者的主要操作是卷积，而后者的主要操作是自注意力机制。CNN 是一种传统的基于卷积操作的神经网络，在传统的计算机视觉任务如图像分类、目标检测等任务中取得了巨大成功。而 ViT 是一种新兴的神经网络模型，它通过自注意力机制来捕捉序列中的长距离依赖关系，取得了超越 CNN 的效果。

卷积神经网络 (CNN) 是一种传统的基于卷积操作的神经网络，在图像分类、目标检测等任务中取得了巨大成功。2012 年，AlexNet[K] 首次将卷积神经网络引入图像分类任务，取得了 ImageNet[D] 比赛的冠军。之后，VGG[S]、GoogLeNet[S]、ResNet[H] 等模型相继提出，不断提高了图像分类任务的性能。这些模型通过卷积、池化、Batch Normalization、激活函数等操作，逐渐提取出输入数据的高层次特征。其中，卷积操作是 CNN 的核心操作，卷积操作的本质是一种特殊的线性变换，它通过卷积核与输入数据的点乘操作，将输入数据的局部特征提取出来。卷积操作的局部性质使得 CNN 能够有效地提取空间特征，因此 CNN 在图像处理领域取得了巨大的成功。

Transformer[V] 是一种在自然语言处理领域取得巨大成功的模型，通过自注意力机制来捕捉序列中的长距离依赖关系。在自然语言处理领域，基于 Transformer 的两种经典架构是 BERT[D] 和 GPT[R]，两者分别通过掩码学习和自回归学习来学习文本的表示。这类模型通过自注意力、Layer Normalization、激活函数等操作，逐渐提取出输入数据的高层次特征，取得了超越传统 RNN 模型的效果。受到自然语言处理领域的

启发，Vision Transformer(ViT) 将图像分割成若干个图块并组成序列，然后按照 Transformer 的方式对序列进行处理。ViT 在图像分类、目标检测等任务中取得了超越 CNN 的效果，表明自注意力机制在计算机视觉领域的有效性。

然而，CNN 和 ViT 是两种截然不同的模型，它们的结构、特征表示、泛化能力等方面存在很大的差异。CNN 是一种基于卷积操作的神经网络，它通过卷积核与输入数据的点乘操作，提取输入数据的局部特征。而 ViT 是一种基于自注意力机制的神经网络，它可以对序列中任意两个元素之间的关系进行建模。这两种模型的差异可能导致它们在不同的任务上表现出不同的优劣势。因此，本文的目标是比较 CNN 和 ViT 的差异，揭示两者的优劣势，探讨两者的内在原因。本文着重关注如下问题：

CNN 和 ViT 存在哪些差异？哪些原因导致了两者的差异？

考虑到两者的差异，本文将从全局和局部特征、降维可视化、模型性能和泛化能力的角度对 CNN 和 ViT 进行比较。通过这些方法，本文将详细对两者关注的特征表示、模型性能和通用性、适用任务等角度进行比较，揭示两者的优劣势。

实验结果表明，CNN 和 ViT 在不同的任务上表现出不同的优劣势。受到归纳偏置的作用，CNN 倾向于学习更规则的区域，其形状类似于卷积核的感受野。而 ViT 则倾向于学习更不规则的区域，呈现成片的形状。这种差异使得 CNN 更适用于学习局部特征，而 ViT 更适用于学习全局特征，并更为偏向于细节特征，同时也更容易过拟合到噪声上。此外，ViT 的较低层就学习到高级语义特征，而 CNN 需要逐层汇总才能学习到高级语义特征。通过降维可视化，本文发现 CNN 在细粒度数据集上具有更好的效果。

为了探究 CNN 和 ViT 对哪些因素敏感，本文选择了不同的数据集和任务进行分析。本文发现，CNN 在纹理、形状等细节特征上相比 ViT 具有更好的性能，同时具备更好的变换不变性。此外，CNN 在跨数据集泛化能力上更优于 ViT，这可能是因为归纳偏置的假设对于所有图像来说都是通用的。

总的来说，本文的主要贡献在于：

(1) 本文揭示了 CNN 和 ViT 存在本质上的差异，并从理论上分析了两者的区别和可能的原因。

(2) 本文提出了基于特征、降维可视化、性能和泛化能力等方面的比较方法，详细比较了 CNN 和 ViT 的优劣势，以及具体的影响因素。

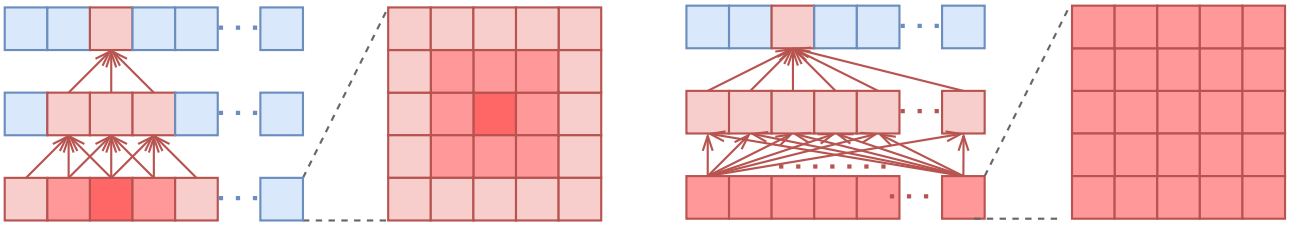
(3) 虽然 ViT 如今在图像分类、目标检测等任务中取得了很好的效果，但本文发现 CNN 在各种任务上仍然具有优势，同时具备更好的细粒度、纹理形状等特征的提取能力和跨数据集的泛化能力，这对于未来的研究具有指导意义。

2 相关工作

本章节将介绍本文的相关工作，包括卷积神经网络、Transformer、Vision Transformer 等模型。

2.1 卷积神经网络

卷积神经网络 (CNN) 是一种经典的视觉模型，在图像分类、目标检测、图像分割等任务中有着大量经典应用。CNN 的历史最早可以追溯到 LeNet-5[L]，这是一种用于手写数字识别的卷积神经网络。之后，AlexNet[K] 在 2012 年 ImageNet 比赛中取得了冠军，超越所有传统的机器学习方法，引发了深度学习的热潮。AlexNet 采用了多层卷积和池化操作，通过堆叠这些操作，逐渐提取出输入数据的高层次特征。之后，VGG[S]、GoogLeNet[S] 等模型相继提出，不断提高了图像分类任务的性能。这些模型通过卷积、池化、Batch



(a) 卷积操作 (b) 自注意力操作

图 1 卷积操作和自注意力操作

Normalization、激活函数等操作，逐渐提取出输入数据的高层次特征，并消除了数据分布的偏差。2015 年，ResNet[H] 提出了残差连接，解决了模型梯度消失和爆炸的的问题，使得模型层数可以进一步加深。DenseNet 也采用了与 ResNet 类似的思路，只不过将残差操作变为了堆叠操作。之后，神经网络架构搜索 (NAS) 技术也逐渐成为研究热点，如 RegNet[R]、EfficientNet[T] 等模型通过自动搜索技术，设计出了更加高效的网络结构。此外，轻量化模型也成为研究热点，如 MobileNet[H]、ShuffleNet[Z] 等模型通过设计轻量化的网络结构，实现了在移动设备上的高效推理。除了图像分类以外，CNN 在目标检测、图像分割、图像生成等任务中也有着大量应用，如 Faster R-CNN[R]、Mask R-CNN[H]、U-Net[R] 等模型，并取得了巨大成功。

2.2 Transformer

Transformer 是一种新兴的神经网络模型，最早由 Vaswani 等人提出[V]，用于自然语言处理任务。传统的循环神经网络 (RNN)[S] 和长短时记忆网络 (LSTM)[G] 等模型在处理长序列数据时存在梯度消失和爆炸的问题，同时也无法并行计算，而 Transformer 解决了上述问题，因此在自然语言处理领域取得了巨大成功。之后，BERT[D]、GPT[R]、T5[R] 等模型相继提出，通过改进预训练方式不断提高自然语言处理任务的性能。受到 Transformer 在自然语言处理领域的启发，一些工作尝试将 Transformer 引入计算机视觉任务[R]，然而并未取得超越 CNN 的性能。直到 Vision Transformer(ViT)[D] 的提出，Transformer 才在计算机视觉领域取得了巨大成功。ViT 将图像分割成若干个图并组成序列，并通过大规模数据集预训练，取得超越 CNN 的性能。之后，DeiT[?]、CaiT[T]、PVT[W] 等模型相继提出，不断提高图像分类任务的性能。此外，Transformer 在目标检测、图像分割、图像生成等任务中也有着大量应用，如 DETR[C]、SegFormer[X] 等模型，并取得了巨大成功。

3 定理

卷积神经网络 (CNN) 和 Vision Transformer(ViT) 是两种截然不同的模型，前者的主要操作是卷积，而后者的主要操作是自注意力机制。本章节将通过理论推导进行分析，比较 CNN 和 ViT 的差异和可能的内在原因。

CNN 是一种特殊的神经网络，经典的 CNN 模型包括 AlexNet、VGG、ResNet 等，这些模型采用卷积、池化、Batch Normalization、激活函数等操作，通过堆叠这些操作，逐渐提取出输入数据的高层次特征。其中，卷积操作是 CNN 的核心操作，卷积操作的本质是一种特殊的线性变换，它通过卷积核与输入数据的点乘操作，将输入数据的局部特征提取出来。卷积操作的局部性质使得 CNN 能够有效地提取空间特征，因此 CNN 在图像处理领域取得了巨大的成功。卷积操作可以表示为公式 1：

$$y_{i,j} = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} x_{i+m,j+n} \cdot w_{m,n}, \tag{1}$$

其中，x 是输入数据，w 是卷积核，y 是输出数据，k 是卷积核的大小。卷积操作的局部性质使得 CNN 能够有效地提取空间特征，但是卷积操作的局限性也很明显，卷积核的大小是固定的，因此 CNN 只能提取固定大小的局部特征，这限制了 CNN 的表达能力。

Vision Transformer 是一种新兴的神经网络模型，它按图像的空间结构将图像分割成若干个图并组成序列，然后通过自注意力、Layer Normalization、激活函数等操作对序列进行处理。ViT 的核心操作是自注意力机制，自注意力机制是一种全局性的操作，它可以对序列中的任意两个元素之间的关系进行建模。自注意力机制的计算公式如下：

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{(W_q Q)(W_k K^T)}{\sqrt{d_k}})(W_v V), \tag{2}$$

其中，Q、K、V 分别是查询、键、值，W_q、W_k、W_v 是权重矩阵，d_k 是维度。自注意力机制首先计算查询 Q 和键 K 之间的相似度，然后通过 Softmax 函数将相似度转换为概率分布，从而建模序列中任意两个元素之间的关系。之后，再通过概率分布对值 V 进行加权求和，得到最终的输出。自注意力机制的全局性质使得 ViT 能够建模序列中任意两个元素之间的关系，因此 ViT 在图像处理领域取得了很好的效果。

从上述公式中，可以看出两种操作的本质区别：

(1) 卷积是一种局部操作，它通过卷积核与输入数据的点乘操作，提取输入数据的局部特征；而自注意力是一种全局操作，它可以对序列中任意两个元素之间的关系进行建模。

(2) 卷积是一种多对一关系，卷积核接收一定范围内的输入，然后输出一个值；而自注意力是一种多对多关系，它接收序列中的所有元素，然后输出一个新的序列。

图 1展示了两种操作的区别。子图 (a) 是卷积操作，每一层的元素只与上一层附近的元素有关，因此卷积操作是一种局部操作；子图 (b) 是自注意力操作，每一层的元素与上一层的所有元素都有关，因此自注意力操作是一种全局操作。由于这一特性，卷积操作形成了对每个元素不同程度的关注，越靠近的元素关注程度越高；而自注意力操作形成了对每个元素相同程度的关注，每个元素都与其他元素有关。卷积操作带来了某种归纳偏置 (Inductive Bias)，可以认为卷积操作基于一种先验假设：

空间上相近的元素之间更具有相关性，空间上不相近的元素之间几乎不具有相关性。

从因果推理的角度，可以对上述假设进行解释。图 2展示了卷积操作和自注意力操作的因果图，其中 X 表示上一层的所有元素，Y 表示下一层的所有元素，T 表示某种映射关系的干预。对于自注意力操作来说，$X \to Y$ 表示自注意力操作，是一个全局映射，而 T 不存在，因此 $X \to Y$ 之间不受任何干预。对于卷积操作来说，则是在 $X \to Y$ 的基础上引入了某种映射关系的干预 T，限制了 X 和 Y 之间的关系，使得全局关系变为局部关系。干预 T 可以形象地解释归纳偏置的作用，它使得空间上相近的元素之间更具有相关性，空间上不相近的元素之间几乎不具有相关性。

综上所述，CNN 和 ViT 的差异主要在于操作的局部性质和全局性质，这一性质可能导致两种模型呈现出不同的现象：

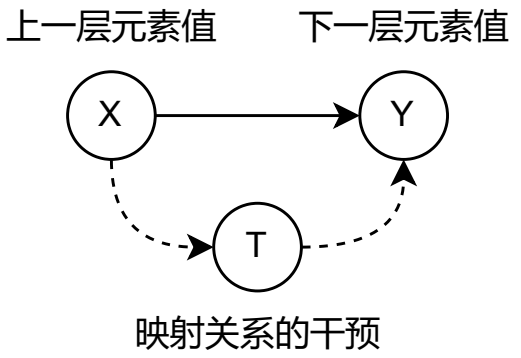


图 2 卷积操作和自注意力操作的因果图

- CNN 更适合建模局部关系，更关注局部细节特征，因为卷积操作是一种局部操作，可以有效地提取输入数据的局部特征，建模元素与附近元素之间的相关性。
- ViT 更适合建模全局关系，更关注整体特征，因为自注意力操作是一种全局操作，可以对序列中任意两个元素之间的关系进行建模。
- 归纳偏置的假设可能导致两者在性能和泛化能力上的差异，这一假设可能会限制模型的性能，因为该假设不一定完全成立；该假设也可能会提高模型的泛化能力，因为在不同的任务下该假设可能是通用的。

在下面的章节中，我们将通过实验验证上述分析，比较 CNN 和 ViT 的特征表示，以及性能和泛化能力，探讨两者的差异和可能的内在原因。

4 方法

本章节将介绍本文使用的一些方法，包括基于全局和局部特征的分析、基于降维可视化的分析、模型性能和泛化性的比较等。

4.1 基于注意力范围分析

Grad-CAM（Gradient-weighted Class Activation Mapping）是一种用于可视化深度学习模型的类激活图（CAM）的技术，它通过利用模型中的梯度信息来生成区分不同类别的图像区域，能够准确地定位模型在做出分类决策时所关注的是图像哪部分区域，这有助于解释模型的预测结果、理解模型的决策过程以及识别模型的弱点。

Grad-CAM 的基本原理和步骤如下：

1. 前向传播首先，将图像输入深度学习模型中进行前向传播，得到模型的输出，此项目中是图像分类任务。

2. 反向传播接着，计算目标类别对于模型最后一层特征图的梯度。其中目标类别是可以进行选择的，可使模型针对想要观察的类别标签进行梯度的计算。

3. 特征图权重计算将目标类别的梯度与模型最后一层特征图进行加权相加，得到每个特征图的权重。其中权重反映了每个特征图对于目标类别的重要性。

4. 生成类激活图将特征图与对应的权重相乘并求和，得到最终的类激活图。这个类激活图表示了模型在做出分类决策时所关注的图像区域。

5. 可视化将类激活图叠加到原始输入图像上，产生彩色的热力图。热力图上的颜色表示了模型在分类决策中所关注的区域。通常，更亮的颜色对应于模型更关注的区域。

Grad-CAM 的优点在于它不需要对模型进行修改，只需在前向传播和反向传播中捕获梯度信息即可。因此，Grad-CAM 适用于各种深度学习模型和任务，并且能够提供直观且解释性强的可视化结果。Grad-CAM 已被广泛应用于图像分类、目标检测、图像分割等领域，以帮助深度学习模型的理解和调试。

4.2 基于全局和局部特征的分析

分析全局和局部特征通常需要对神经网络的隐藏层输出特征进行定量的计算。而分析神经网络的 (隐藏) 层表示往往具有挑战性，因为它们的特征分布在大量神经元上。这种分布式方面也使得跨神经网络的表示难以进行有意义的比较。中心内核对齐 (CKA) 解决了这些挑战，实现了网络内部和网络之间表征的定量比较。通常 CKA 算法将 $x \in \mathbb{R}^{m \times p_1}$ 以及 $y \in \mathbb{R}^{m \times p_2}$ 作为输入，其中 m, p_1, p_2 分别表示样本的个数，以及两层网络神经元的个数。定义 $K = XX^T$ 以及 $L = YY^T$ 作为两个网络层的 $Gram$ 矩阵，CKA 计算如下：

$$CKA(K, L) = \frac{HSIC(K, L)}{\sqrt{HSIC(K, K)HSIC(L, L)}}$$

其中 $HSIC$ 是希尔伯特-施密特独立准则。CKA 对表示的正交变换 (包括神经元的置换) 是不变性的，其归一化项保证了对各向同性尺度的不变性。这些特性使得神经网络隐藏表征的比较和分析变得有意义。

同时为了去研究两种模型 ViT 和 $ConvNeXt$ 之间关于高低频信息的处理能力，我们通过傅里叶变换对图像特征图进行分析。图像频域分析是一种将图像从空间域（空间坐标）转换到频域（频率坐标）的方法。在频域中，图像表示为频谱的形式，反映了图像中不同频率的分量。傅立叶变换是一种常用的频域分析方法，可以将图像从空间域转换到频率域。傅立叶变换将图像表示为频率分量的幅度和相位信息，使得图像在频域中的特征更加明显。傅立叶变换在图像处理中被广泛应用，如图像压缩、去卷积、特征提取等。对于一个大小为 $M \times N$ 的图像傅里叶变换公式如下：

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-i2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)}$$

其中 $f(x, y)$ 表示图像在空间域中的像素强度，$F(u, v)$ 表示图像在频率域中的频率分量，u 和 v 是频率索引。傅里叶变换将图像从空间域转换到频率域，使得图像在频域中的特征更加明显。通过对图像的频域分析，可以更好地理解图像的特征和结构，为图像处理和分析提供更多的信息。

4.3 基于降维可视化的分析

降维可视化技术是理解和比较高维数据特征的一种有效工具，特别是在模型特征表示上的理解。通过将高维特征空间映射到二维或三维空间，可以观察和比较卷积神经网络（CNN）和视觉变压器（ViT）模型提取的特征之间的差异。本文使用了一种流行的非线性降维技术：t-Distributed Stochastic Neighbor Embedding（t-SNE）。

t-SNE 是一种高级降维技术，与 PCA 相比，它更擅长保留数据的局部结构，并在低维空间中展示数据点之间的相对关系。在使用 t-SNE 进行降维之前，通常会进行数据的标准化处理，以确保不同的特征尺度不会影响降维的结果。t-SNE 的核心思想是在低维空间中寻找一个数据点的分布，这个分布可以最好地反映高维空间中数据点之间的相似性。

4.4 模型性能和泛化性的比较

在传统图像分类任务上，在 ImageNet 数据集上测试准确率是评估模型性能的主要方法，自从 2012 年 AlexNet 首次在 ImageNet 上取得突破性进展以来，ImageNet 准确率一直是评估模型性能的重要指标。然而，随着大模型的出现和数据集的增大，ImageNet 准确率已经不能完全反映模型的性能。首先，在大模型选择在更大、更复杂的数据集上进行预训练之后，ImageNet 已经是一个相对简单的数据集，其类别数量和复杂程度已经不能满足真实世界的需求。其次，ImageNet 准确率不能反映模型的泛化能力，即模型在未见过的数据上的表现。

为了探究模型性能受哪些因素影响，本文选择了 ImageNet-X 和 PUG-ImageNet 2 种数据集，这两个数据集提供了丰富的属性注释，可以通过控制变量法来定量研究模型对于哪些属性更加敏感。ImageNet-X 数据集是 ImageNet 数据集的扩展，其中包含了一些常见的数据扰动，如模糊、噪声、对比度变化等。PUG-ImageNet 数据集是一个人为构造的数据集，其中包含了一些特定属性，如颜色、形状、纹理等。通过这两个数据集，本文可以研究模型对于不同属性的敏感性，从而揭示模型的内在特性。

为了探究模型的泛化性能，本文选择了跨域、跨数据集和图像变换等任务。跨域任务是指模型在一个数据集上训练，然后在另一个数据集上测试，这可以评估模型对于不同数据分布的适应能力。跨数据集任务是指模型在一个数据集上训练，然后在另一个数据集上测试，这可以评估模型对于不同数据集的泛化能力。图像变换任务是指模型在一个数据集上训练，然后在另一个数据集上测试，这可以评估模型应对图像变换或破坏时的适应能力。通过这些任务，本文可以研究模型的泛化能力，从而揭示模型的通用性。

5 实验

本章节将介绍本次实验的具体细节，包括实验环境、数据集、评价指标、实验结果等。之后将展示实验结果并进行分析，包括全局与局部特征分析、可视化分析、性能与泛化能力对比等。最后本章节将进行总结。

5.1 实现细节

本文的实验代码使用 Python 语言编写，选用 ConvNeXt-Base 和 ViT-16/B 两种模型进行对比，两种模型同样在 ImageNet-21K 上进行预训练，选用 224x224 的图像分辨率。此外，两种模型具备相似的参数量，前者为 89M，后者为 87M。本文的所有实验在一台拥有 4 块 NVIDIA V100 GPU 的服务器上进行，实验使用 PyTorch 框架进行实现。

5.2 数据集与评价指标

本文用于降维可视化的主要数据集有 CIFAR-10 和 CUB200 两个数据集。CIFAR-10 数据集包含 10 个不同的类别，如汽车、鸟类和猫等，这些类别之间的差异较大。CUB200 数据集专注于鸟类图像，包含 200 个子类别，这些子类别间的差异相对较小，主要体现在鸟的种类、羽毛颜色和姿态等细微特征上。这两个数据集的选择旨在考察模型在不同数据集上的性能差异，以及对于不同类别之间的差异性的处理能力。

本文选择 ImageNet-X 和 PUG-Imagenet 两种数据集对模型的影响因素进行评估。ImageNet-X 数据集是一个用于探索模型错误影响因素，数据集提供了关于 16 种变化因素的详细人机注释，例如姿态、风格等。这使得可以有针对性地分析模型的错误类型。而 PUG-ImageNet 数据集是一个合成数据集，该数据集中的图像使用软件引擎生成，允许系统地改变每个对象的姿势、大小、纹理、光线和背景等因素，从而通过

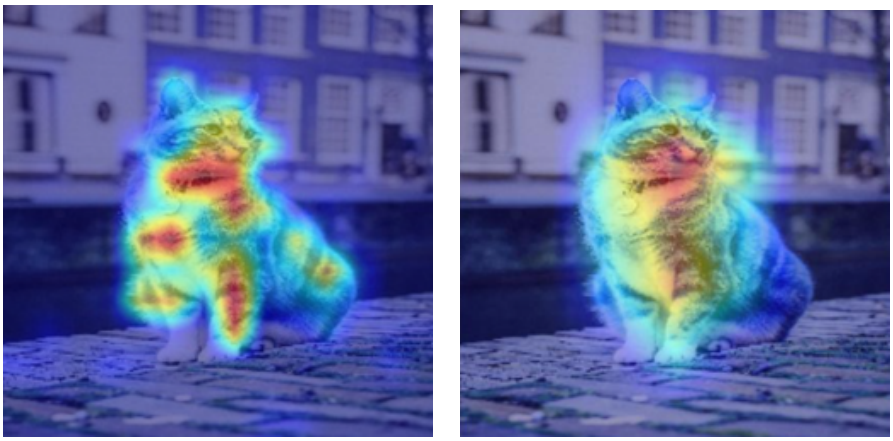
控制这些因素来评估模型的性能。

为了探究模型的跨域泛化能力，本文选择 ImageNet-Sktech、ImageNet-V2、ImageNet-R、ImageNet-A 4 种数据集，这些数据集都是 ImageNet 数据集的变种，用于评估模型在不同数据集上的泛化能力。ImageNet-Sktech 数据集是一个手绘图像数据集，ImageNet-V2 数据集是一个用于评估模型在真实世界中的泛化能力的数据集，ImageNet-R 数据集是一个用于评估模型在不同分辨率下的泛化能力的数据集，ImageNet-A 数据集是一个用于评估模型在不同摄像机角度下的泛化能力的数据集。此外，为了探究模型的跨数据集泛化能力，本文选择 Pets、Caltech101、CIFAR-100、DTD、Flowers102、SUN397、EuroSAT 7 种数据集，这些数据集拥有不同的类别，如动物、植物、环境、纹理、卫星图像等，用于评估模型在不同数据集上的泛化能力。

本文的主要评价指标有轮廓系数、域内准确率和域外准确率等，这些指标用于评估模型在不同数据集上的性能差异。轮廓系数是一个衡量聚类质量的指标，其值越接近于 1，说明聚类内的相似度高而聚类间的差异大，聚类效果越好。域内准确率和域外准确率分别用于评估模型性能，模型通过在域内、域外数据集上进行微调之后进行测试得到准确率，通过比较模型在域内、域外数据集上的准确率，可以评估模型的泛化能力。

5.3 实验结果

5.3.1 全局和局部特征分析



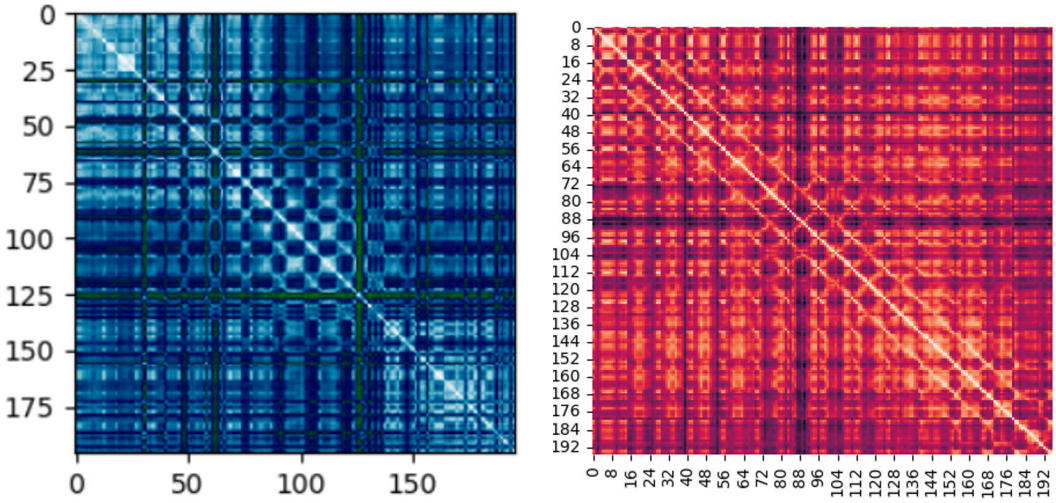
(a) ViT (b) ConvNeXt

图 3 ViT 和 ConvNeXt 的 Grad-CAM 可视化注意力图

通过分别选取 ViT 模型和 ConvNeXt 模型的图像编码器的最后一层，进行梯度计算，得到图 3的结果。在图中可以较为明显的看到，两种模型在进行图像类时都能将注意力放在前景物体上，ViT 的关注区间比较发散，而 ConvNeXt 比较集中，呈现从关注中心向四周散开的高斯形状。

这种观察表明 ViT 和 ConvNeXt 在处理图像时可能存在不同的注意力分布模式。ViT 显示出的较为发散的关注区间可能反映了其自注意力机制的特性，该机制允许模型在图像的不同区域之间建立长距离的依赖关系。因此，ViT 在处理图像时可能能够建立全局的信息，从而导致关注区间相对发散。相比之下，ConvNeXt 呈现出的关注区间更加集中，呈现出从猫头为中心的高斯形状。这可能反映了传统卷积神经网络的特点，即它们会受到归纳偏置的影响。在处理图像时可能更加注重临近范围特定目标的捕捉，因此其

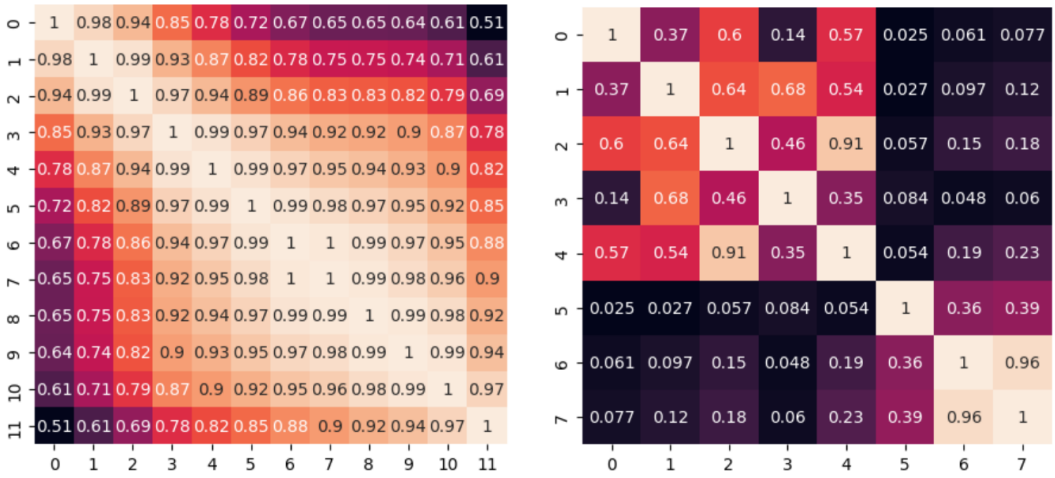
关注区间更加集中于目标的中心区域。从上述表现中可初步说明 ViT 模型的注意力范围更远。



(a) ViT (b) ConvNeXt

图 4 ViT 和 ConvNeXt 局部特征之间的相似度

为了进一步验证上述采样，本文对两个网络的局部相似度进行可视化分析。通过计算模型局部特征块两两之间的余弦相似度，得到图 4的结果。如图所示，ConvNeXt 的局部相似度矩阵中呈现出多条颜色较亮的斜平行线，这说明其中的每个局部块特征都与上下左右的相邻局部块特征之间存在更大的相似性。相比之下，ViT 模型的局部块之间并没有存在明显的空间关系限制。这说明 ConvNeXt 更关注空间相似性，即空间关系会影响图像表征；ViT 能够在全局范围内捕捉图像的长距离依赖关系，不受到像素之间空间关系的影响。这也说明了 ConvNeXt 形成类似高斯形状的注意力图，而 ViT 更为发散的原因。



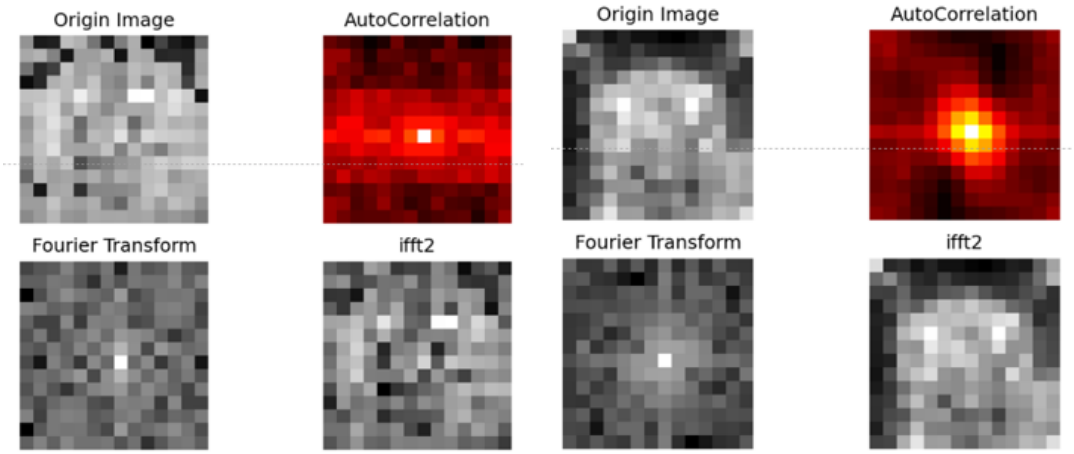
(a) ViT (b) ConvNeXt

图 5 ViT 和 ConvNeXt 层之间的 CKA 相似度

为了对比 ConvNeXt 和 ViT 模型在特征理解能力上的区别，本文提取模型逐层的特征进行分析。ViT 注意力层输出特征，ConvNeXt 使用卷积层输出特征。本文通过 CKA 方法分别计算两种模型每层结构的

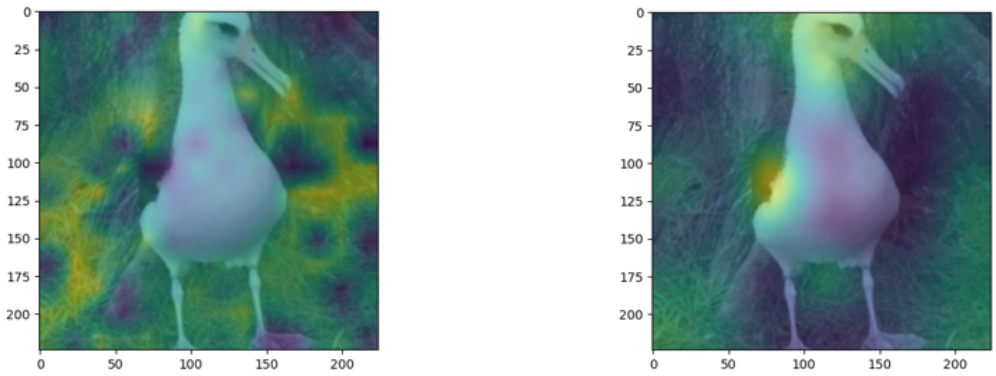
特征相似度，CKA 方法可以有效度量不同维度向量之间的相似度，从而应对层之间的维度变化。得到层级之间的 CKA 相似度矩阵如图 5，从图中可看出，ViT 模型在浅层时特征和高层时特征已有比较明显的相似性，说明其能在浅层时就学到比较高级的语义信息；相比而言，ConvNeXt 模型的浅层和深层特征之间的相似度具有较大的区别，说明 CNN 需要逐层卷积才能逐渐聚合高层语义特征。

从归纳偏置的视角来看，ViT 由于不受任何约束，浅层模块就可以学习长程注意力，提取全局特征。然而，在 CNN 的卷积操作中，每个区域只能与周围区域进行局部交互，这限制了浅层模块的全局特征提取能力，只能学习纹理、形状等低级语义信息。直到深层卷积聚合了浅层的局部信息之后，才能学习高级语义信息。这种机制更类似于人类大脑的工作机制，首先提取局部信息，之后聚合全局信息。



(a) ViT (b) ConvNeXt

图 6 ViT 和 ConvNeXt 的频域分析



(a) ViT 特征图 (b) ConvNeXt 特征图

图 7 ViT 和 ConvNeXt 特征图可视化

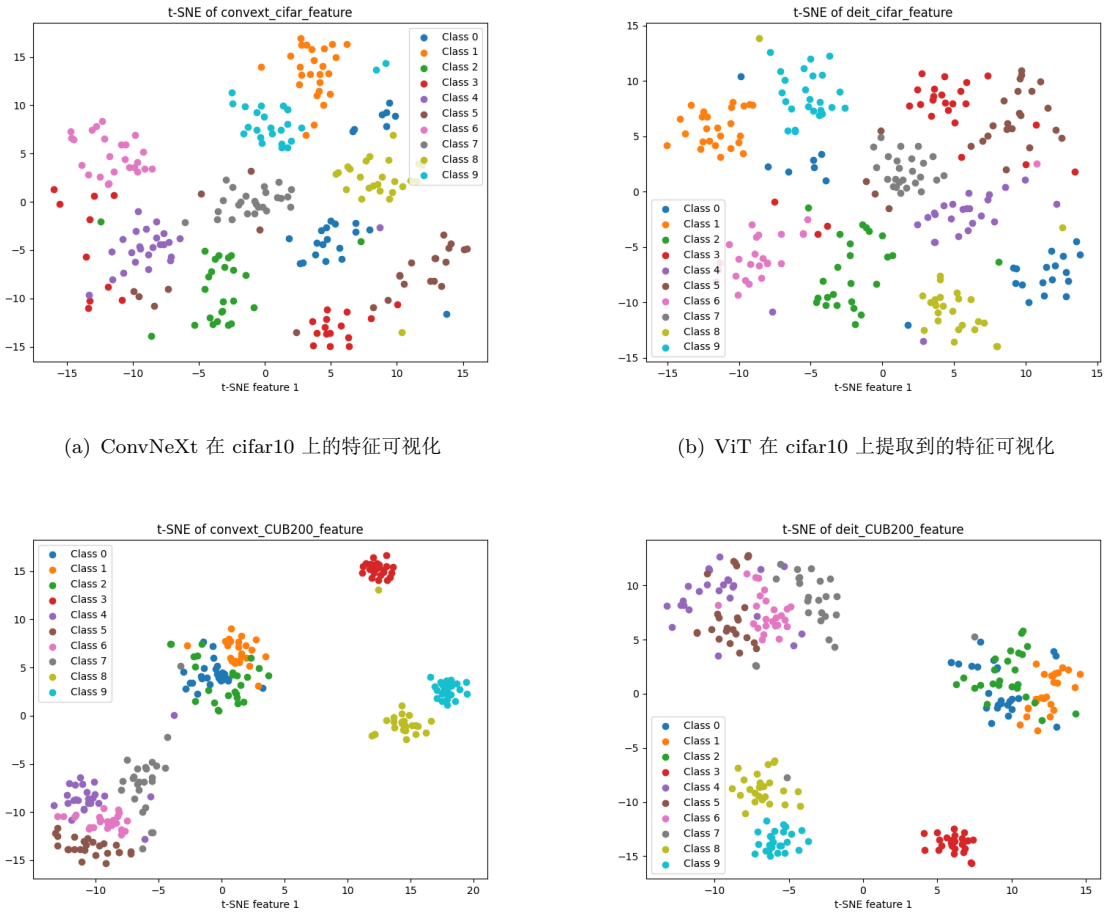
为了进一步说明 ViT 模型与 ConvNeXt 在学习能力上差异，本文对特征进行频域分析。如图 6所示，ViT 的低频和高频信息都受到高度激活，这说明 ViT 能够兼顾低频和高频信息。另外，ConvNeXt 的低频信息激活明显，而高频信息激活并不明显，这说明 CNN 倾向于学习低频信息。通过频域分析，本文发现

ViT 比 CNN 更倾向于细节信息，这也可能导致模型在噪声上容易过拟合。

图 7展示了 ViT 的一种过拟合情形，通过对特征在通道上求均值，得到模型的感兴趣区域。从图中可以看出，ViT 倾向于关注图像中的背景噪声，而 ConvNeXt 是比较精准的关注在了前景物体的关键部位。这可能是由于 CNN 模型的归纳偏置限制了感兴趣区域的形状，使得模型倾向于关注到与任务相关的主要特征。而 ViT 缺乏这种限制，使得模型发散到无关特征上。因此，CNN 的归纳偏置能够在一定程度上能够减轻无关背景细节的影响，减轻过拟合现象。

5.3.2 降维特征分析

在使用 t-SNE 进行降维分析的过程中，在图 8的子图 (a)(b) 可以清楚地看到 CIFAR-10 和 CUB200 两个数据集中不同模型的表现差异。对于通用数据集 CIFAR-10 来说，两种模型都有不错的特征分布；然而对于细粒度数据集 CUB200 来说，ConvNeXt 的特征分布要比 ViT 更优。



(a) ConvNeXt 在 cifar10 上的特征可视化

(b) ViT 在 cifar10 上提取到的特征可视化

(c) ConvNeXt 在 CUB200 上的特征可视化

(d) ViT 在 CUB200 上的特征可视化

图 8 ViT 和 ConvNeXt 在 CUB200 和 cifar10 上的特征可视化

为了量化特征分布的优劣程度，本文通过计算轮廓系数，进一步验证了这些观察结果。表 1反映了两种模型的轮廓系数。从表中可以看出，在 CIFAR-10 中 ConvNeXt 的轮廓系数和 ViT 接近，表现都非常好。在 CUB200 数据集上，ConvNeXt 的轮廓系数明显高于 ViT。这强调了 ConvNeXt 在处理包含细微局部差

异的数据集时的优势，尤其是在 CUB200 这样的高度专业化的数据集上。ConvNeXt 的高性能主要归因于其优秀的局部特征提取能力，这使得它在相似类别之间也能够有效地区分。

表 1 ViT 和 ConvNeXt 在两种数据集上的轮廓系数

	ViT	ConvNeXt
CUB200	0.113	0.130
CIFAR10	0.117	0.124

5.3.3 性能与泛化能力对比

要想分析图像分类错误的影响因素，仅仅识别出错误的物体类别远远不够，关键在于找出这些错误的具体原因。例如，某些模型可能对数据分布的某些方面特别敏感，比如纹理变化。在这种情况下，当物体的纹理与它所训练的数据不同的时候，模型可能会一直出错。识别错误类型可以让有针对性地收集和重新训练数据，比黑盒方法具有优势。ImageNet-X 数据集提供了关于 16 种变化因素的详细人机注释，例如姿态、风格等。本文选用该数据集，通过控制变量对影响因素进行分析，从而有针对性地分析模型的错误类型。

为了分析每种因素对模型错误率等影响，本文提出如下指标进行评估：

$$errorratio(factor) = \frac{1 - accuracy(factor)}{1 - accuracy(overall)} \tag{3}$$

其中 $accuracy(overall)$ 是 ImageNet-1K 验证的总体准确率，$accuracy(factor)$ 是仅调整因素 $factor$ 的图像上的准确率。该指标可以衡量模型在给定因素上的性能相对于其整体性能的影响因子。

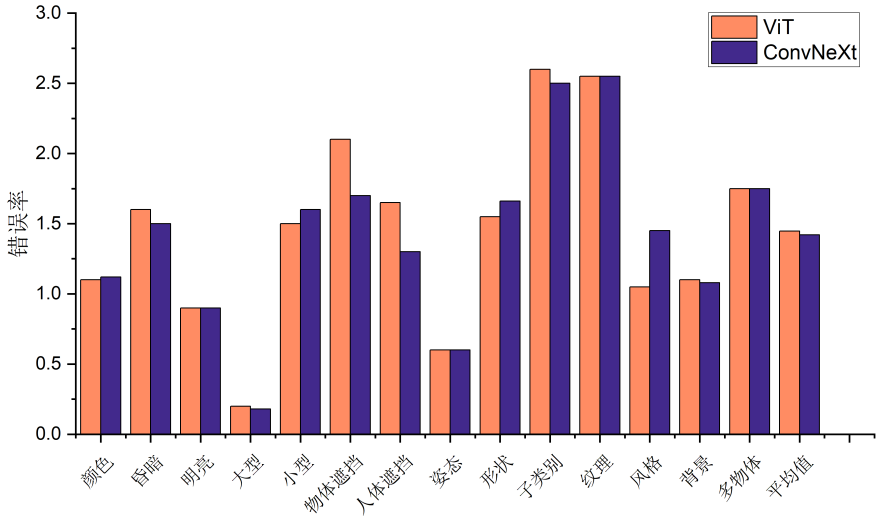


图 9 ViT 和 ConvNeXt 在不同因素上的错误率

图 9中展示了两种模型对于各种因素的错误率。结果表明两种模型在风格、遮挡、小型物体上的错误率具有较大差异，其中 ViT 对风格更不敏感，而 ConvNeXt 对物体尺寸、遮挡更不敏感。此外，两种模型在纹理、子类别、多物体、昏暗环境中都容易犯错误，而在明亮、大型物体、不同姿态的环境中表现良好。

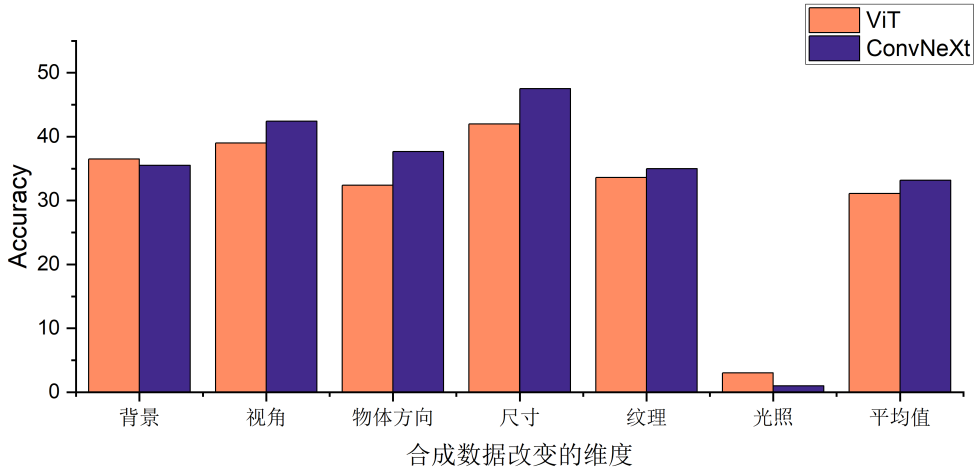


图 10 合成数据准确性比较

与人类注释的数据不同，合成数据集允许精确控制相机角度、物体位置和纹理等因素。PUG-ImageNet 是一个合成数据集，包含 ImageNet 类别的照片般逼真的图像并提供属性标签。该图像使用软件引擎生成，允许系统地改变每个对象的姿势、大小、纹理、光线和背景等因素。本文测试了 PUG-ImageNet 中六个不同因素的 Top-1 准确性结果，如图 10所示，ConvNeXt 在视角、物体方向、尺寸、纹理等因素上相比 ViT 都具备更好的性能，而 ViT 仅在背景、光照因素上具备优势。这说明 CNN 学习的特征对于物体形变来说更为有效。

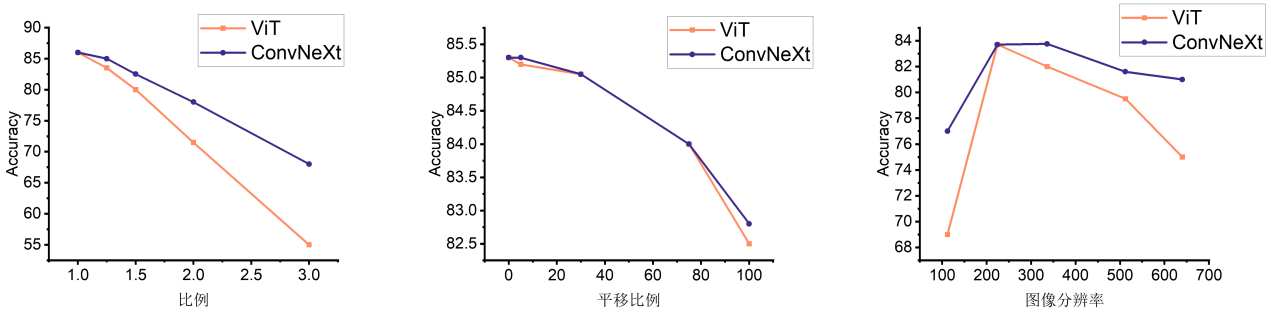


图 11 变换不变性实验结果

在实际场景中，数据可能会经过各种各样的变换，如缩放、平移、分辨率改变等，对数据不变具备不变性的模型可以更好应对真实场景的数据变换。以往的实验[A]证明神经网络的性能在简单的输入数据变换下（例如，将图像向左或向上平移几个像素）可能高度不稳定。本文通过对图片进行缩放、平移和分辨率改变对模型的变换不变性进行评估，结果如图 11所示。从图中可以看出，ConvNeXt 在不同变换因素下的准确率始终优于 ViT。此外，ConvNeXt 在图像分辨率变换上的表现稳定，而在缩放、平移变换上出现了较大的性能下降，这说明 CNN 具备更强大的变换不变性，同时也仍然具备改进空间。

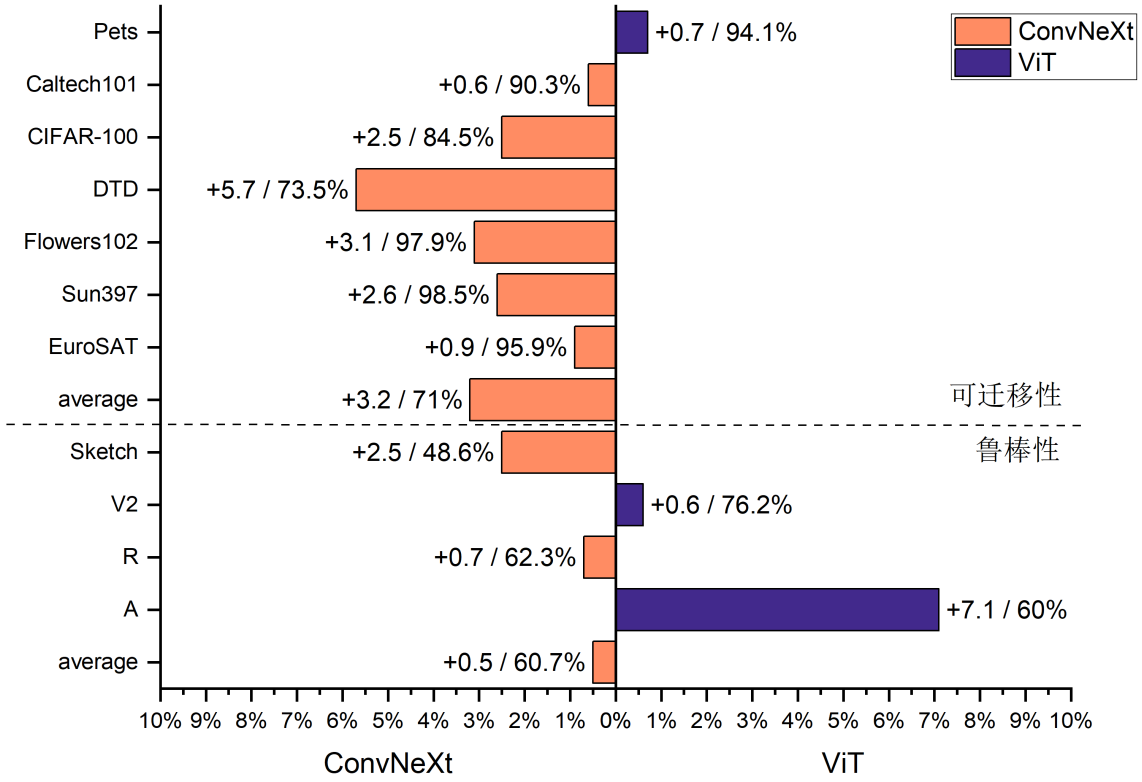


图 12 鲁棒性和可迁移性的比较

模型性能不仅限于在训练同分布数据集上的准确率，有的模型可能在训练分布的数据上表现出色，但很难将这种表现推广到数据分布的转变。这些转变可以由自然扰动引起，如大气条件（例如，雾、雨），相机噪声或物体位置和方向的变化。模型鲁棒性衡量模型在数据分布变化方面的适应能力，本文在几个包含许多不同类型的自然变化和损坏的数据集测试上评估了模型的鲁棒性，包括 ImageNet-V2，ImageNet-A，ImageNet-R，ImageNet-Sketch，这些数据集对 ImageNet 的图像风格进行处理，包括素描、卡通等不同风格以及对抗性图像，通过对这些图像进行评估，可以反映模型鲁棒性。

模型的迁移学习性能表明其适应新任务的能力，良好的可转移性允许模型使用最小的代价进行快速微调。模型在不显著降低性能的情况下适应这些转变的能力是一种有价值的度量标准，可以衡量其效用和泛化能力。本文使用在 ImageNet 上预训练的模型，其主要包含自然图像。之后本文采用了 7 个不同的数据集，分别是 EuroSAT，Sun397，Flowers102，DTD，CIFAR-100，Caltech101，Pets，包括如动物、植物、纹理、卫星图像等不同场景。通过对这些数据集进行评估，可以反映模型的泛化能力。

如图 12所示，两种模型在鲁棒性上的表现相似，其中 ViT 在对抗图像上的表现更好，ConvNeXt 则在素描图像上的表现更好，这表明 ConvNeXt 可能更擅长提取边缘和轮廓信息。此外，ConvNeXt 几乎在所有跨数据集任务上的性能更好，这表明 CNN 的归纳偏置可能是所有数据集之间通用的假设，使得模型具备更为强大的跨数据集迁移学习能力。

5.4 总结

本章节通过对 ViT 和 ConvNeXt 模型的全局和局部特征分析、降维特征分析、性能与泛化能力对比等方面的实验，对两种模型的优劣势进行了详细的分析。通过对两种模型的注意力图、局部特征相似度、层级特征相似度、频域分析、特征图可视化等方面的对比，本文发现 ViT 模型更倾向于全局特征提取，而 ConvNeXt 模型更倾向于局部特征提取。通过对两种模型在 CIFAR-10 和 CUB200 数据集上的特征可视化和轮廓系数的对比，本文发现 ConvNeXt 模型在处理细粒度数据集上的性能更优。通过在 ImageNet-X 和 PUG-ImageNet 数据集上的性能对比以及变换不变性测试，本文发现 ConvNeXt 模型对于形状、纹理等因素更不敏感。此外，通过跨域和跨数据集迁移学习实验，本文发现 ConvNeXt 模型具备更强大的泛化能力。

综上所述，ViT 模型在全局特征提取、对抗性图像学习等方面具备优势。而 ConvNeXt 模型在局部特征提取、细粒度数据集、跨数据集迁移学习等方面具备优势，此外总体性能也优于 ViT。因此，CNN 可以更适合于处理细粒度数据集、跨数据集迁移学习、噪声学习等任务，而 ViT 则更适合于全局特征提取、对抗性图像学习等任务。

6 结论

卷积神经网络 (CNN) 和 Vision Transformer(ViT) 是两种截然不同的模型，前者的主要操作是卷积，而后者的主要操作是自注意力机制。本章节将通过理论推导进行分析，比较 CNN 和 ViT 的差异和可能的内在原因。之后通过对全局和局部特征、降维可视化、模型性能和泛化性能的比较，进一步分析和比较了 CNN 和 ViT 的差异和造成差异的内在原因。通过实验，本文得出了以下主要结论：

(1) 受到卷积结构的限制，CNN 倾向于关注固定形状的特征区域，且需要从浅到深逐步提取特征，这与人类大脑的工作机制更为接近。而 ViT 倾向于学习更自由的特征区域，能够在浅层就学到高级语义信息。

(2) CNN 对于形状、纹理、尺寸、视角等因素具备更强的鲁棒性，同时对图像变换具备更强大不变性。

(3) CNN 具备更强大的跨数据集迁移学习能力，而 ViT 在对抗性图像上表现更好。

本文的工作揭开了 CNN 和 ViT 的内在差异，为未来的工作提供了新的启发。本文认为虽然 ViT 在图像分类任务上取得了很好的效果，但是 CNN 可能会以一种工具的形式继续存在，例如 DeTR 等模型的结构中也包含了 CNN 的部分结构，此外 CNN 也可以用来学习位置编码等信息。因此，未来的工作可以通过结合 CNN 和 ViT 的优势，设计更加强大的模型，提高模型的泛化能力和鲁棒性。

参考文献

[Li et al.(2022)Li, Liu, Yang, Peng, and Zhou] LI Z, LIU F, YANG W, et al. A survey of convolutional neural networks: Analysis, applications, and prospects[J/OL]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 6999-7019. DOI: 10.1109/TNNLS.2021.3084827.

[Dosovitskiy et al.(2021)Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, and Mostafa Dehghani] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[A]. 2021. arXiv: 2010.11929.

[Krizhevsky et al.(2017)Krizhevsky, Sutskever, and Hinton] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J/OL]. Commun. ACM, 2017, 60(6): 84–90. https://doi.org/10.1145/3065386.

[Deng et al.(2009)Deng, Dong, Socher, Li, Li, and Fei-Fei] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C/OL]//2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255. DOI: 10.1109/CVPR.2009.5206848.

[Simonyan et al.(2015)Simonyan and Zisserman] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[A]. 2015. arXiv: 1409.1556.

[Szegedy et al.(2015)Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke, and Rabinovich] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C/OL]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 1-9. DOI: 10.1109/CVPR.2015.7298594.

[He et al.(2015)He, Zhang, Ren, and Sun] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [J/OL]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 770-778. https://api.semanticscholar.org/CorpusID:206594692.

[Vaswani et al.(2017)Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 6000–6010.

[Devlin et al.(2019)Devlin, Chang, Lee, and Toutanova] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C/OL]//BURSTEIN J, DORAN C, SOLORIO T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186. https://aclanthology.org/N19-1423. DOI: 10.18653/v1/N19-1423.

[Radford et al.(2018)Radford and Narasimhan] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[C/OL]//2018. https://api.semanticscholar.org/CorpusID:49313245.

[Lecun et al.(1998)Lecun, Bottou, Bengio, and Haffner] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J/OL]. Proceedings of the IEEE, 1998, 86(11): 2278-2324. DOI: 10.1109/5.726791.

[Radosavovic et al.(2020)Radosavovic, Kosaraju, Girshick, He, and Dollár] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R, et al. Designing network design spaces[C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 10425-10433. DOI: 10.1109/CVPR42600.2020.01044.

[Tan et al.(2019)Tan and Le] TAN M, LE Q V. Efficientnet: Rethinking model scaling for convolutional neural networks [A]. 2019. arXiv: 1905.11946.

[Howard et al.(2017)Howard, Zhu, Chen, Kalenichenko, Wang, Weyand, Andreetto, and Adam] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[A]. 2017. arXiv: 1704.04861.

[Zhang et al.(2017)Zhang, Zhou, Lin, and Sun] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[A]. 2017. arXiv: 1707.01083.

[Ren et al.(2017)Ren, He, Girshick, and Sun] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149. DOI: 10.1109/TPAMI.2016.2577031.

[He et al.(2017)He, Gkioxari, Dollár, and Girshick] HE K, GKIOXARI G, DOLLáR P, et al. Mask r-cnn[A]. 2017. arXiv: 1703.06870.

[Ronneberger et al.(2015)Ronneberger, Fischer, and Brox] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//NAVAB N, HORNEGGER J, WELLS W M, et al. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241.

[Schmidt(2019)] SCHMIDT R M. Recurrent neural networks (rnns): A gentle introduction and overview[A]. 2019. arXiv: 1912.05911.

[Graves(2012)] GRAVES A. Long short-term memory[M/OL]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 37-45. https://doi.org/10.1007/978-3-642-24797-2_4.

[Raffel et al.(2023)Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[A]. 2023. arXiv: 1910.10683.

[Ramachandran et al.(2019)Ramachandran, Parmar, Vaswani, Bello, Levskaya, and Shlens] RAMACHANDRAN P, PARMAR N, VASWANI A, et al. Stand-alone self-attention in vision models[M]. Red Hook, NY, USA: Curran Associates Inc., 2019.

[Touvron et al.(2021)Touvron, Cord, Douze, Massa, Sablayrolles, and Jegou] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers and distillation through attention[C/OL]//MEILA M, ZHANG T. Proceedings of Machine Learning Research: Vol. 139 Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021: 10347-10357. https://proceedings.mlr.press/v139/touvron21a.html.

[Wang et al.(2021)Wang, Xie, Li, Fan, Song, Liang, Lu, Luo, and Shao] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C/OL]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 548-558. DOI: 10.1109/ICCV48922.2021.00061.

[Carion et al.(2020)Carion, Massa, Synnaeve, Usunier, Kirillov, and Zagoruyko] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C/OL]//Computer Vision –ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I. Berlin, Heidelberg: Springer-Verlag, 2020: 213–229. https://doi.org/10.1007/978-3-030-58452-8_13.

[Xie et al.(2021)Xie, Wang, Yu, Anandkumar, Alvarez, and Luo] XIE E, WANG W, YU Z, et al. Segformer: Simple and efficient design for semantic segmentation with transformers[C/OL]//RANZATO M, BEYGELZIMER A, DAUPHIN Y, et al. Advances in Neural Information Processing Systems: Vol. 34. Curran Associates, Inc., 2021: 12077-12090. https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

[Azulay et al.(2019)Azulay and Weiss] AZULAY A, WEISS Y. Why do deep convolutional networks generalize so poorly to small image transformations?[J]. Journal of Machine Learning Research, 2019, 20(184): 1-25.