Transformer 在计算机视觉领域的应用

作　　者： **郎志远**

学科专业： **计算机科学与技术**

上海大学材料基因组工程研究院学院

2024 年 01 月

摘　要

Transformer 模型，最初在自然语言处理（NLP）领域取得了革命性的成功，它以其独特的自注意力机制和并行处理能力，成功地解决了长距离依赖问题。这种架构的灵活性和强大的特征表示能力引起了计算机视觉研究者的兴趣，他们开始探索将 Transformer 应用于视觉任务，以期利用其在处理序列数据方面的优势。在视觉基准测试中，基于 Transformer 的模型已经证明了其在图像分类、目标检测、语义分割等任务中的有效性，甚至在某些情况下超越了传统的卷积神经网络和循环神经网络。这些模型的成功展示了 Transformer 在处理视觉数据时的潜力，尤其是在捕捉图像的全局上下文信息方面。此外，Transformer 模型在视觉任务中的泛化能力较强，对视觉特定归纳偏置的需求相对较少，这使得它们在多样化的视觉任务中具有广泛的应用前景。本文对视觉 Transformer 模型进行了细致的分类，深入分析了它们在高级视觉处理任务、基础视觉任务以及其他计算视觉任务中的应用，最后本文还探讨了 Transformer 在视觉领域所面临的挑战，并提出了未来研究的可能方向。

关键词：Transformer；计算机视觉；神经网络；

一、引言

深度学习网络已成为现代人工智能技术的核心，它们根据不同的应用场景采用不同的网络结构。例如，全连接网络通过叠加线性和非线性层来处理数据[1, 2]。为了处理图像等具有平移不变性的数据，卷积神经网络（CNN）引入了卷积和池化层[3, 4]。递归神经网络则通过循环单元来分析序列化数据或时间序列[5, 6]。Transformer 架构是一种新兴的神经网络，它通过自注意力机制来捕捉数据的内在特性，并在人工智能领域展现出广泛的应用前景[7]。

在自然语言处理领域，Transformer 架构已经实现了显著的进步[8]。Vaswani 等人[9]首次提出了基于注意力机制的 Transformer 模型，用于机器翻译和句子结构分析。Devlin 等人[10]开发了 BERT 模型，它在处理语言时考虑了单词的双向上下文，从而在多个 NLP 任务上达到了当时的最佳性能。Brown 等人[11]则在大规模文本数据上训练了 GPT-3，这是一个拥有 1750 亿参数的 Transformer 模型，它在多种下游 NLP 任务中展现出无需微调的强大能力。这些基于 Transformer 的模型在 NLP 领域取得了重大突破[12],[13]。受到这些成就的激励，研究者们开始探索将 Transformer 应用于计算机视觉任务。尽管 CNN 在视觉任务中占据主导地位，但 Transformer 已经显示出其作为潜在替代方案的潜力。例如，Mark Chen 等人[14]训练了一个序列 Transformer 来预测像素，其在图像分类任务中的表现与 CNN 相当。ViT 模型则直接将 Transformer 应用于图像块，实现了对整个图像的分类。Dosovitskiy 等人[15]提出的 ViT 模型在多个图像识别任务中达到了最先进的性能。此外，Transformer 还被应用于目标检测、语义分割、图像处理和视频理解等多个视觉领域。鉴于其在视觉任务中的优异表现，越来越多的研究致力于开发基于 Transformer 的模型。随着基于 Transformer 的视觉模型不断涌现，保持对最新进展的了解变得日益困难。因此，本文旨在综述视觉 Transformer 的最新发展，并探讨未来可能的研究方向。

二、Transformer 基本结构

在 2017 年，Vaswani 及其团队首次介绍了 Transformer 模型，该模型由六个编码器-解码器对构成（如图 1 所示），每个编码器包含一个多头自注意力机制和一个前馈神经网络。解码器部分则由三层组成，其中第一层和最后一层与编码器结构相似，而中间层则采用了交叉注意力机制，其关键输入（键 K 和值 V）源自编码器的输出。本节将详细阐述 Transformer 架构中各个组件的功能和特性。

2.1 自注意力机制

在自注意力机制中，输入数据首先被转化为三个关键向量：查询（Q）、键（K）和值（V），它们都具有相同的维度。这些向量是通过与特定的权重矩阵（W^Q、W^K、W^V）相乘得到的。接下来，注意力权重的计算遵循以下步骤：

步骤 1：使用 S = Q * KT 计算不同输入向量之间的分数；

步骤 2：为了增强梯度稳定性，对分数进行归一化， $S_n = S/\sqrt{d_k}$；

步骤 3：使用 softmax 函数将分数转换为概率，P = softmax(S_n)

步骤 4：通过 V * P 获得加权值矩阵，Z = V * P。

这个过程的核心思想在于，通过计算输入向量之间的分数，模型能够确定在处理当前元素时应该给予其他元素多少注意力。归一化步骤有助于稳定梯度，而 softmax 函数则将分数转换为概率，使得模型能够更有效地分配注意力。在解码器中，编码器-解码器注意力层与编码器中的自注意力层类似，但解码器使用来自编码器的 K 和 V，而 Q 则来自前一层。为了解决自注意力层无法直接捕捉位置信息的问题，Transformer 引入了位置编码。这通过在输入嵌入中添加一个与模型维度相同的额外维度来实现。位置编码使用正弦数使得模型能够通过相对位置信息来学习注意力，从而处理更长的序列。

多头注意力机制进一步扩展了自注意力，通过将输入数据分割成多个头，每个头关注不同的信息。这样，模型可以在不同的子空间中并行处理信息，增强了对输入数据多角度关联的学习。多头注意力保持了参数总量不变，但通过分散注意力，提高了模型的表达能力。

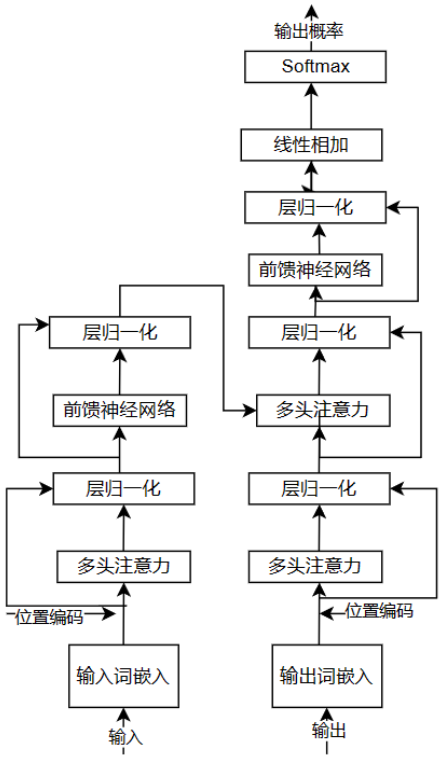

图 1 Transformer 的网络架构

2.2 前馈神经网络及其层归一化

基本数据探索我们主要关注缺失值：训练集的标签属性列中，基础的房屋信息如各种在 Transformer 架构中，除了自注意力子层，编码器和解码器的每个模块还包含一个全连接的前馈神经网络。这个网络由两个线性变换组成，中间夹着一个 ReLU 激活函数。前馈网络对序列中的每个元素应用相同的变换，尽管在不同的层之间使用了不同的权重。

具体的计算过程可以如下：序列首先通过第一个线性层，然后通过 ReLU 激活函数，激活后的输出再经过第二个线性层，得到最终的前馈网络输出。

这种设计允许网络在保持位置不变性的同时[16]，对序列中的每个元素进行非线性变换，增强了模型的表达能力。通过在不同层使用不同的参数，网络能够学习到序列中元素之间的复杂关系。为了维持数据特征分布的一致性，Transformer 在注意力和前馈网络层之间引入了层归一化。然而，残差连接的直接合并可能导致随着网络层数的增加，主分支的激活值逐渐累积，振幅增大，这可能引起训练过程中的不稳定性。为了解决这一问题，Swin-Transformer v2[17]采用了一种改进的归一化策略，将层归一化层移至每个子层之后。这样，每个残差块的输出在与主分支合并之前都会经过归一化处理。这种后归一化方法有效避免了激活振幅的累积，使得网络在增加层数时保持更稳定的激活幅度，从而显著提升了大型视觉模型的训练稳定性。

三、视觉 Transformer

3.1 高/中层视觉任务

DETR（Detection Transformer）是一种基于 Transformer 架构的新型目标检测框架，其网络结构如图 2 所示，旨在实现从目标检测到检测结果的端到端流程。尽管 DETR 在理论上具有潜力，但它在实际应用中面临训练时间长和对小目标检测效果不佳的挑战。为了克服这些问题，朱熹周等人[18]提出了一种改进的 DETR 版本，即可变形 DETR，该方法通过专注于参考点周围的有限关键区域，而非全局特征图，有效减少了计算负担并加快了训练收敛速度。这种设计还使得模型能够轻松地融合多尺度特征，从而在保持较低推理成本的同时，实现了比原始 DETR 更高的检测性能。此外，孙罗伟[19]等人分析了 DETR 收敛缓慢的原因，并发现这主要与 Transformer 解码器中的交叉注意力机制有关。他们提出了一种仅使用编码器的 DETR 变体，该变体在保持较高检测精度的同时，显著缩短了训练时间。同时，他们还提出了 TSP-FCOS 和 TSP-RCNN 两种基于 Transformer 的集合预测模型，这些模型通过特征金字塔技术优化了仅编码器 DETR，高鹏等人则提出了空间调制共注意力机制，该机制进一步加速了 DETR 的收敛过程。将 SMCA 集成到 DETR 中，可以在不牺牲太多推理效率的情况下，显著减少训练周期，实现了与 DETR 相当的平均精度，这些改进使得 DETR 在实际应用中更加可行，尤其是在处理小目标和复杂场景时。

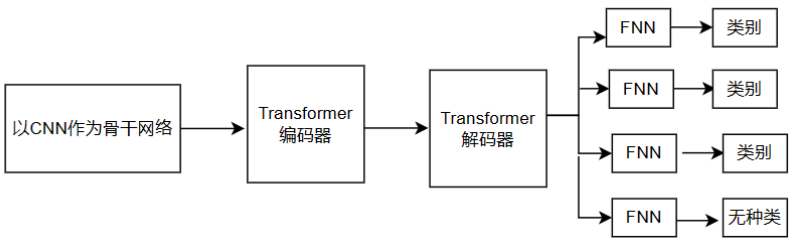

图 2 DERT 的网络架构

3.2 低层视觉任务

在视觉处理领域，Transformer 模型的应用主要集中在高级任务如图像分类、分割和检测。然而，对于低级视觉任务，如图像超分辨率和生成，这类研究相对较少。这些任务要求模型生成高质量的图像作为输出[21]，这在技术上比高级视觉任务更具挑战性。陈汉庭等人[22]提出了一种名为图像处理 Transformer（IPT）的模型，它通过利用大规模预训练数据集，有效地发挥了 Transformer 在图像处理方面的优势。IPT 在超分辨率、去噪和去雨等任务上展现了卓越的性能。

IPT 的结构设计包括多个注意力头、编码器、解码器以及多个任务特定的输出部分。这种设计允许模型针对不同的图像处理任务进行优化。在预训练阶段，IPT 使用 ImageNet 数据集，通过引入噪声、雨迹或下采样等方法，生成低质量的图像作为训练数据。这些损坏的图像作为输入，而高质量的原始图像作为目标输出。此外，IPT 还采用了自监督学习方法，以增强模型的泛化能力。经过预训练后，IPT 可以通过微调来适应特定的图像处理任务。在图像去噪任务中，IPT 实现了 2dB 的性能提升，这证明了 Transformer 模型在低级视觉任务中的潜力。

3.3 其他视觉任务

鉴于 Transformer 在自然语言处理（NLP）领域的卓越表现，研究者们开始探索其在多模态任务中的应用潜力，如视频-文本、图像-文本和音频-文本的结合。CLIP（Contrastive Language-Image Pre-training）[23]是一个创新的多模态模型，它通过自然语言作为监督信号，来提升图像表示的学习效率。CLIP 模型结合了一个文本编码器和一个图像编码器，共同训练以预测匹配的文本-图像对。文本编码器基于标准的 Transformer 架构，具备掩蔽自注意力机制，以保持语言模型的预训练优势。图像编码器则可以选择 ResNet 或视觉 Transformer 架构。CLIP 在由互联网收集的 4 亿对图像-文本对上进行训练，这些对通过学习最大化匹配对的余弦相似度，同时最小化错误匹配对的相似度来优化。

CLIP 模型能够根据文本描述来识别和匹配相应的图像，而 DALL-E[24]则进一步发展了这一概念，它能够基于文本描述生成全新的图像。DALL-E 是一个拥有 120 亿参数的多模态 Transformer 模型，它在包含 330 万文本-图像对的数据集上进行了自回归训练。训练过程分为两个阶段：首先，使用离散变分自编码器将高分辨率的 256x256 RGB 图像压缩成 32x32 的图像标记；接着，在第二阶段，通过训练一个自回归 Transformer 来学习图像和文本标记之间的联合分布。DALL-E 能够从零开始创造各种风格的图像，如逼真照片、卡通画和表情符号，甚至能够在保持文本描述一致性的同时对现有图像进行扩展。此外，丁明等人[25]开发的 CogView 模型也采用了类似的 VQ-VAE 标记器，但特别支持中文文本输入。他们声称 CogView 在性能上超越了 DALL-E 和基于生成对抗网络（GAN）的方法，并且 CogView 不需要额外的 CLIP 模型来对 Transformer 生成的样本进行排序。

四、总结与展望

Transformer 因其在计算机视觉领域的出色表现和巨大潜力，正逐渐成为研究的焦点。近年来，已经有一些方法被提出来利用 Transformer 在视觉任务中的应用，这些方法在多种视觉任务上取得了显著成果，包括构建骨干网络[26],[27],[28]、处理高级和低级视觉任务以及视频分析。尽管如此，Transformer 在视觉领域的潜力还有待进一步挖掘，存在一些挑战需要克服。

目前，尽管已经有许多基于 Transformer 的模型被提出来解决视觉问题，但这些尝试仍处于初步阶段，有待进一步优化。例如，ViT 模型虽然遵循了 NLP 领域的 Transformer 架构，但针对计算机视觉的专门优化版本尚需开发。此外，Transformer 在视觉任务中的泛化能力和鲁棒性仍然是一个挑战，因为与 CNN 相比，它缺乏一些归纳偏置[29]，且对大规模数据集的依赖性较强。尽管 ViT 在图像分类任务上表现优异，但在目标检测等任务上，其效果并未超越 CNN。为了使预训练的 Transformer 更好地适应多样化的视觉任务，还需要更多的研究。此外，Transformer 的鲁棒性也是一个关键问题，尽管已有研究，但仍未完全解决。Transformer 模型的参数量巨大，这可能影响其在视觉任务中的可解释性。开发高效的 Transformer 模型以适应计算机视觉的需求仍然是一个挑战。例如，基本的 ViT 模型在处理图像时的计算成本非常高，而相比之下，轻量级的 CNN 模型如 GhostNet[30]在性能相近的情况下，计算成本要低得多。虽然已有一些 Transformer 压缩方法，但它们在视觉领域的适用性仍有待验证。因此，迫切需要开发出既高效又适用于视觉任务的 Transformer 模型，以便在资源受限的设备上部署。

参考文献

[1] F. Rosenblatt, The Perceptron, a Perceiving and Recognizing Automaton Project Para. Buffalo, New York, USA: Cornell Aeronautical Lab., 1957.

[2] J. Orbach, “Principles of neurodynamics. perceptrons and the theory of brain mechanisms,” Arch. General Psychiatry, vol. 7, no. 3, pp. 218–219, 1962.

[3] Y. LeCun et al., “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp 2278–2324, 1998.

[4] A. Krizhevsky et al., “ImageNet classification with deep convolutional neural networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Parallel distributed processing: Explorations in the microstructure of cognition,” vol. 1, pp. 318–362, 1986.

[6] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp 1735–1780, 1997.

[7] D. Bahdanau et al., “Neural machine translation by jointly learning to align and translate,” in Proc. Int. Conf. Learn. Representations, 2015.

[8] A. Parikh et al., “A decomposable attention model for natural language inference,” in Proc. Empir. Methods Natural Lang. Process., 2016, pp. 2249–2255.

[9] A. Vaswani et al., “Attention is all you need,” in Proc. Conf. Neural Informat. Process. Syst., 2017, pp. 6000–6010.

[10]J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language

understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol., 2019, pp. 4171–4186.
[11]T. B. Brown et al., "Language models are few-shot learners," in Proc. Conf. Neural Informat. Process. Syst., 2020, pp. 1877–1901.
[12]K. He et al., "Deep residual learning for image recognition," in Proc. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[13]S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," in Proc. Conf. Neural Informat. Process. Syst., 2015, pp. 91–99.
[14]M. Chen et al., "Generative pretraining from pixels," in Proc. Int. Conf. Mach. Learn., 2020, pp. 1691–1703.
[15]A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Representations, 2021.
[16] N. Carion et al., "End-to-end object detection with transformers," in Proc. Eur. Conf. Comput. Vis., 2020, pp. 213–229.
[17]X. Zhu et al., "Deformable DETR: Deformable transformers for endto-end object detection," in Proc. Int. Conf. Learn. Representations, 2021.
[18]S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in Proc. Conf. Comput. Vis. Pattern Recognit., 2021, pp. 6881–6890.
[19]H. Chen et al., "Pre-trained image processing transformer," in Proc. Conf. Comput. Vis. Pattern Recognit., 2021, pp. 12299–12310.
[20]L. Zhou et al., "End-to-end dense video captioning with masked transformer," in Proc. Conf. Comput. Vis. Pattern Recognit., pp. 8739–8748, 2018.
[21]S. Ullman et al., High-Level Vision: Object Recognition and Visual Cognition, Vol. 2. Cambridge, MA, USA: MIT press, 1996.
[22]R. Kimchi et al., Perceptual Organization in Vision: Behavioral and Neural Perspectives. Hove, East Sussex, U.K.: Psychology Press, 2003.
[23]J. Zhu et al., "Top-down saliency detection via contextual pooling, J. Signal Process. Syst., vol. 74, no. 1, pp 33–46, 2014.
[24]J. Long et al., "Fully convolutional networks for semantic segmentation," in Proc. Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3431–3440.
[25]H. Wang et al., "Max-deeplab: End-to-end panoptic segmentation with mask transformers," in Proc. Conf. Comput. Vis. Pattern Recognit., 2021, pp. 5463–5474.
[26] R. B. Fisher, "CVonline: The evolving, distributed, non-proprietary, on-line compendium of computer vision," 2008. Accessed: Jan. 28, 2006. [Online]. Available: http://homepages.inf.ed.ac. uk/rbf/CVonline
[27]N. Parmar et al., "Image transformer," in Proc. Int. Conf. Mach. Learn., 2018, pp. 4055–4064.
[28]Y. Zeng et al., "Learning joint spatial-temporal transformations for video inpainting," in Proc. Eur. Conf. Comput. Vis., 2020, pp. 528–543.
[29]K. Han et al., "Transformer in transformer," in Proc. Conf. Neural Informat. Process. Syst., 2021. [Online]. Available: https:// proceedings.neurips.cc/
[30]H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," 2021, arXiv:2105.05537.