2D Transformer

2D Transformer是一种基于 Transformer架构的神经网络，专门用于处理二维数据，如图像。 Transformer 最初是在自然语言处理（ NLP）领域提出的，用于处理序列数据。然而，由于其强大的自注意力机制（ Self-Attention Mechanism）， Transformer架构已经被扩展到了计算机视觉和其他领域。以下是 2D Transformer的一些主流技术和方法：

1. ViT (Vision Transformer)：
ViT是Google Research提出的一种将 Transformer
应用于图像分类的方法。它将图像分割成多个 patches，然后将这些 patches作为序列输入到
Transformer中。 ViT展示了 Transformer在图像识别任务上的有效性，并激发了后续许多基于
Transformer的图像模型的研究。
2. Swin Transformer：
Swin Transformer是一种采用层次化分块（ Hierarchical Patching）策略的 Transformer
模型。它将图像分成不同大小的块，并通过局部和全局自注意力机制来捕捉图像的不同层次的
特征。 Swin Transformer在多个视觉任务上取得了优异的性能。
3. DeiT (Data-Efficient Image Transformer)：
DeiT
强调了数据效率，通过使用少量数据进行训练来减少过拟合，并利用知识蒸馏技术从大型预训
练模型中提取知识。 DeiT证明了 Transformer可以在较小的数据集上也能取得很好的性能。
4. ConViT (Convolutional Vision Transformer)：
ConViT尝试将传统的卷积神经网络（ CNN）和 Transformer
结合起来。它使用卷积层来提取图像的局部特征，然后使用 Transformer
来捕捉全局依赖关系。这种方法结合了 CNN的空间局部性优势和 Transformer
的长距离依赖捕捉能力。
5. Twins (Transformer in Transformer)：
Twins通过在 Transformer的每个层中嵌入一个小型的 Transformer（称为 Twin Transformer
）来增强模型的特征表示能力。这种方法使得模型能够在不同的尺度上捕捉图像特征，并提高
了模型的性能。
6. CvT (Convolutional Vision Transformer)：
CvT将卷积操作融入到 Transformer
的自注意力机制中，使得模型可以在进行自注意力计算之前先进行特征的空间过滤。这种方法
旨在结合 CNN和Transformer的优点。

这些技术和方法的共同点是它们都利用了 Transformer
的自注意力机制来处理图像数据，但是它们在实现细节上有所不同，包括如何处理图像数据、如何组织
Transformer层、以及如何捕捉不同尺度的特征等。随着研究的深入，我们可以预见将会有更多创新的 2D Transformer架构出现，并在计算机视觉领域取得更多的突破。

Mamba Vision Mamba

https://zhuanlan.zhihu.com/p/680130826

https://arxiv.org/pdf/2401.09417.pdf

ViTamin

https://beckschen.github.io/vitamin

Revision #2
Created 2025-01-11 09:46:28 UTC by Colin
Updated 2026-04-29 07:34:10 UTC by Colin