软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.000000]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

Transformer 的研究及其应用

马燕[1]

[1](河海大学 计算机与信息学院,江苏 南京 210094)

摘 要: Transformer 模型抛弃了传统的 CNN 和 RNN，整个网络结构完全由 Attention 机制组成，该方法不仅在 NLP 领域在计算视觉领域都取得了显著的成就，本文从注意力机制引起，详细阐述 Transformer 的原理和数学推导，并且结合近期的基于 Transformer 在二维视觉和三维视觉的研究，讨论了 Transformer 在计算视觉的优势、特点及其可行性。

关键词: Transformer; 注意力机制; 图像分类; 点云配准
中图法分类号: TP311

中文引用格式: 马燕.Transformer 的研究及其应用.软件学报,2021,32(7). http://www.jos.org.cn/1000-9825/0000.htm
英文引用格式: Ma Y. Research on Transformer and its applications. Ruan Jian Xue Bao/Journal of Software, 2021 (in Chinese). http://www.jos.org.cn/1000-9825/0000.htm

Research on Transformer and its applications

Ma Yan[1]

[1](College of Computer and Information, Hohai University, Nanjing 210094, China)

Abstract: Transformer model abandons the traditional CNN and RNN, the entire network structure is completely composed of Attention mechanism, this method has not only made significant achievements in the field of NLP in the field of computational vision, this paper from the attention mechanism, elaborate the principle and mathematical derivation of Transformer, and combined with recent research based on Transformer in two-dimensional vision and three-dimensional vision, The advantages, features and feasibility of Transformer in computational vision are discussed.

Key words: Transformer; self-attention; image classification; point cloud registration

Transformer 最早在 NLP 领域提出的一种基于注意力机制的模型[1]，在其提出后，在计算视觉领域也有很好的运用。

本文第 1 节介绍 Attnention 注意力机制第 2 节 Transformer 模型的结构，包括编码器、解码器和最后的输出层。第 3 节介绍以 ViT 为例 Transformer 在二维计算视觉分类的应用。第 4 节介绍了以 GeoTransformer 为例子的三维视觉中点云配准任务的应用。最后总结全文。

1 Attention 注意力机制

Attention 机制是 Transformer 模型的基础，事实上 Transformer 的开山之作的的标题即为“Attention Is All You Need”，在了解 Transformer 之前需要先了解注意力机制。

1.1 注意力机制

Attention 是注意力的意思，字面来看，借鉴了人类的注意力机制。

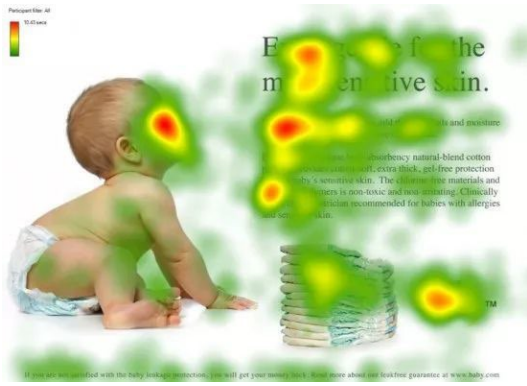

图 1　视觉注意力机制示意图

最为典型的注意力机制是人类的视觉注意力机制，人类视觉可以通过快速全局的图像，获得需要重点关注的最为感兴趣的目标区域，然后对这个区域进行更为细节的观测，获得更为详细的信息。这种注意力机制也就是说让人类能够从全幅的大量信息精选出值得关注的部分细节信息，更够提高视觉信息处理的效率。

深度学习的注意力信息也和人类选择的视觉注意力信息类似[2]，本质上是从众多信息中选择出较为关键的信息。

1.2 Attention的本质思想

Attention 能够从众多信息中找到更为有价值和重要的部分，例如 LSTM 解决了 nlp 中序列长距离的依赖问题，但是词汇超过 200 的时候就会失效，通过 ttention 机制可以较好解决这样的问题。

首先注意力模型从大量信息的 *values* 中筛选出重要的信息，这种重要的信息也是相对于 *query* 而言，比如在上文的婴儿图，*query* 就是观察图像的人，也就是说一个注意力模型，有一个 *query* 和一组 *values*， 然后 *query* 从 *values* 里面筛选出重要的信息

通过 *query* 从这个 *values* 中筛选出重要的信息，也就是计算出 *query* 和 *values* 中信息的相关程度

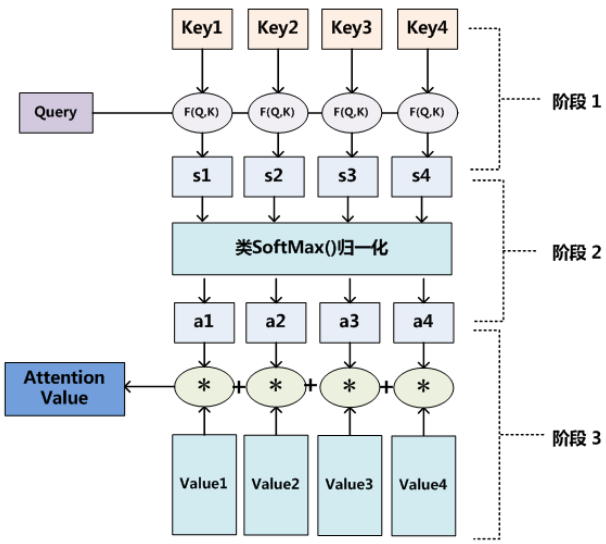

图 2　注意力机制示意图

具体来说 Attention 可以由上图所示，将 *Query*(*Q*) 和 *key* – *value* 键值对映射到输出上，其中每一个 *query*，

key， *value* 输出是所有 *values* 的加权，其中权重是 *Query*(*Q*) 和每个 *key* 计算出。

第一步：计算比较 Q 和 K 的相似度，表示为

$$f(Q,K_i) \quad i=1,2,\cdots,m \tag{1}$$

一般计算方法有点乘（Transformer 使用）：

$$f(Q,K_i)=Q^T K_i \tag{2}$$

第二步：将得到的相似度进行 *SoftMax* 操作，进行归一化：

$$\alpha_i = SoftMax(\frac{f(Q,K_i)}{\sqrt{d_k}}) \tag{3}$$

其中，分母为 $\sqrt{d_k}$ 的原因是为了控制 f 的方差为 1

第三步：针对计算出的权重 α_i 对 V 中所有的 *values* 进行加权求和计算，得到 Attention 向量:

$$attention=\sum_{i=1}^{m}\alpha_i V_i \tag{4}$$

1.3 Self -Attention自注意力机制

自注意力机制是注意力机制中的一种，Self-Attention 模型的架构如下图所示

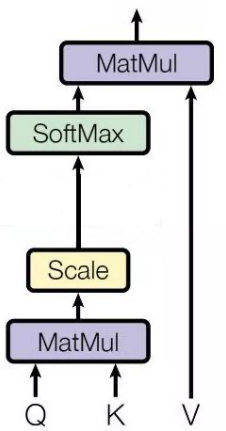

图 3 Self-Attention 自注意力模型示意图

可以看到 Self-Attention 有个三个输入，Q、K、V，这里的三个输入都来自于句子的词向量的线性转化，也就是自注意力机制的“自”指的是同一输入映射到不同的空间，自己产生三个输入。后续的处理和注意力机制相同，减少了对外部信息的依赖。

1.4 Mutli-head Self-Attention多头自注意力模型

实际 Transformer 的实现中是以多头自注意力机制为主，主要是把 Self-Attention 注意力 Z 进行切分成 n 个，然后通过全连接层获得新的 Z'。

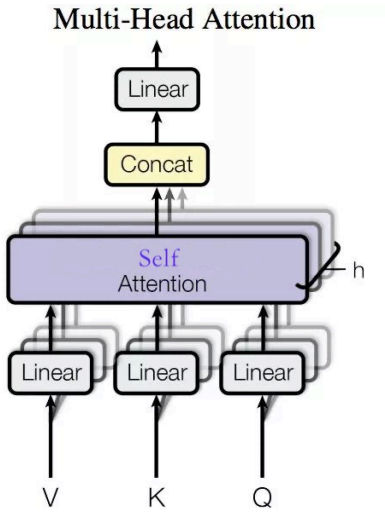

图 4 Mutli-head Self-Attention 模型示意图

也就是说多头把原始信息 source 放入了多个子空间，捕捉到了多个信息，多头保证了可以 attention 主要到不同子空间的信息，捕获到更为丰富的特征信息，在实际 Transformer 的提出中作者发现这样能够达到较好的效果。

2 Transformer

Transformer 实际上可以看成前面所说的 Self-Attention 模型的叠加。

2.1 Transformer结构

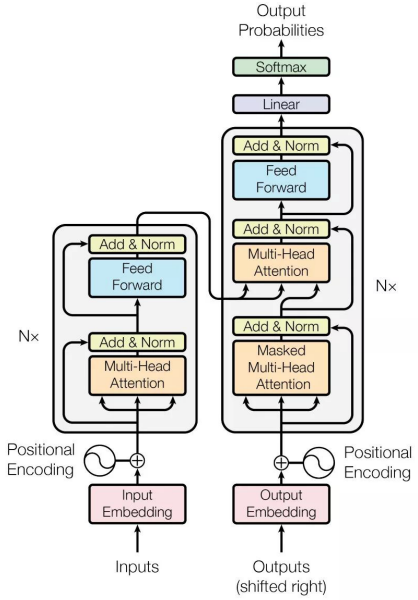

图 5 Transformer 模型结构

简单的讲 Transformer 也是一个 Seq2Seq 模型[3]（Encoder-Decoder 框架的模型），左边一个 Encoders 把输入读进去，右边一个 Decoders 得到输出，如下所示：

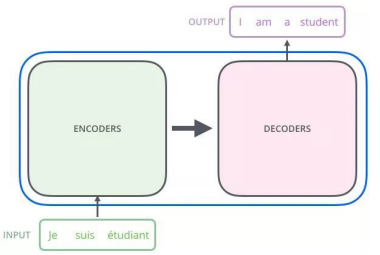

图 6 Transformer 的 Seq2Seq 模型结构

随之产生的问题是 Encoder 和 Decoder 是如何结合的，实际上里面有 n 层的 Encoder，Encoder 的输出会和 Decoder 进行结合。

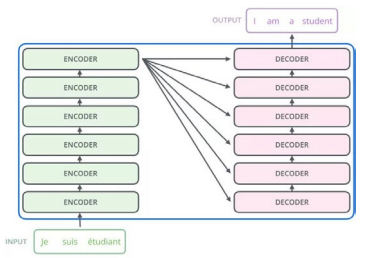

图 7 Transformer 的多层 Encoder-Decoder 模型结构

2.2 Encoder编码器

具体的某一层如图所示

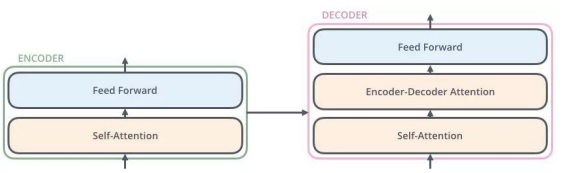

图 7 某一层 Encoder-Decoder 结构

Encoder 是 6 层， 每层包括 2 个 sub-layers 子层：

第一个 sub-layer 子层是上文提到的 Multi-head Self-Attention 多头自注意力层。

第二个 sub-layer 子层是简单的 Feed Forward 前馈神经网络[4]。

其中每一个 sub-layer 都模拟了残差网络，每个 sub-layer 的输出都是：

$$LayerNorm(x + Sub\_layer(x)) \tag{5}$$

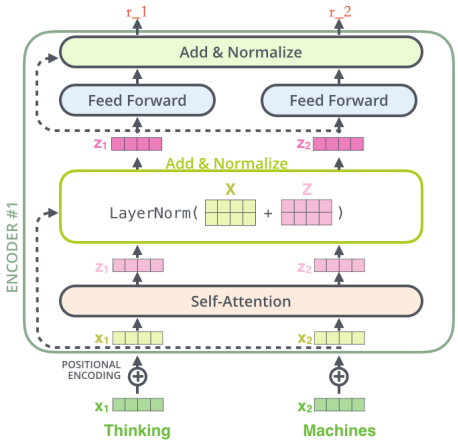

图 8　编码器 Encoder 子层结构

计算步骤如图 8 所示

第一步：深绿色的 x_1 表示 Embedding 层的输出，加上代表 Positional Embedding 的特征向量，得到输入 Encoder 中的特征向量。

第二步：特征向量 x_1 经过 Self-Attention 层变成 z_1，其中残差结构[5]下需要先用 x_1 和 z_1 相加。

第三步：z_1 经过 Feed Forward 层，经过残差结构和自身相加，再讲过 Normalize 层，得到输出向量 r_1。

第四步：因为有多层 Encoder，r_1 作为下一层的输入，循环得到最后一层的结果。

2.3　Decoder解码器

Transformer 的 Decoder 解码器包括 3 个 sub-layer：

第一个 sub-layer 是 Masked Multi-head Self-attention 经过 Mask 操作的多头自注意力层 也是计算的输入的 Attention。

第二个 sub-layer 是 Encoder-Decoder attention 计算，对 Encoder 的输入和 Decoder 的 Masked multi-head self-attention 的输出进行 Attention 计算

第三个 sub-layer 是前反馈神经网络，和 Encoder 编码器中的相同。

2.4　Transformer输出结果

经过 Encoder 编码器和 Decoder 解码器的两大模块，最后的工作是将结果输出。

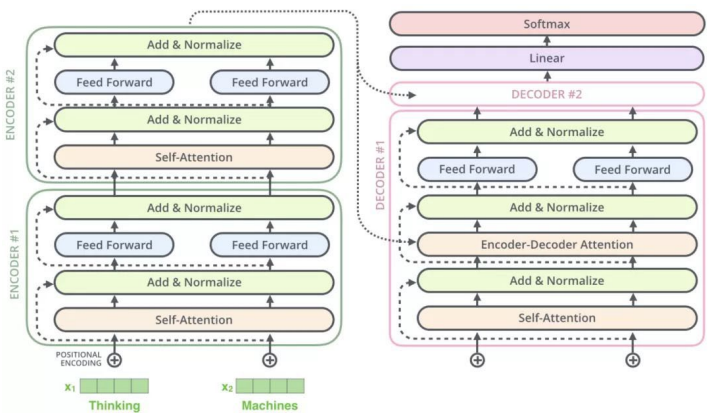

图 8　Transformer 输出示意图

从图中可以看到 Transformer 最后的工作是让解码器的结果通过 Linear 线性层后接一个 softmax，即可得到最终的输出。

3 Transformer 在二维视觉的应用-以 Vision Transformer 为例

在上文 Transformer 的介绍中不难发现，Transformer 和相关的注意力机制最早提出是用于 NLP 领域，使得模型可以并行计算，极大提高了效率，同时模型表现上有一定的提升。但是随着 Transformer 模型的发展，就像视觉注意力机制那样，研究者自然而然地尝试把 Trasnformer 运用到计算视觉领域。但是 Transformer 的输入是 1D 序列，计算视觉中的输入一般是 2D 图像甚至是 3D 视觉，如何处理视觉信息编码序列更好的输入到 Transformer 网络中是 Transformer 在计算视觉领域应用关键问题。

ViT 是 2020 年 Goole 团队提出的将 Transformer 应用在图像分类的模型[6]，模型相对简单效果好，可扩展性强，成为了 Transformer 在计算视觉领域较为关键的研究。

3.1 ViT的结构

针对图像的输入，Vit 将其分成多个 16 16 的 patch，再将每个 patch 投影为固定长度的向量送入 Transformer，后续的 Encoder 操作和原始 Transformer 模型中的完全相同，其中由于要进行图片分类任务，在输入序列中加入了特殊的 token，该 token 的输出为最后的类别预测。

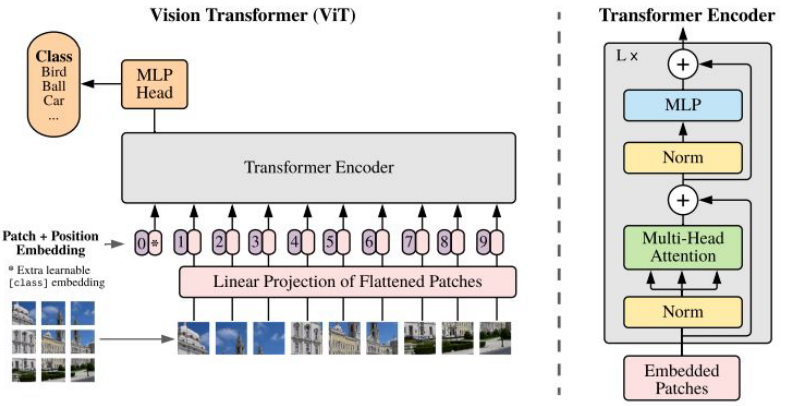

图 9 ViT 模型结构

值得注意的是，Vit 并没有用到 Transformer 的 Decoder 因为 token 可以直接输出最后的预测结果，具体来说一个 ViT Block 可以分为以下步骤：

第一步：patch embedding 把 2 维视觉问题转换成了 seq2seq 问题。

第二步：positional encoding, Vit 同样需要加入位置编码，位置编码可以理解为一张表，表一共有 N 行，N 的大小和输入序列长度相同，每一行代表一个向量，向量的维度和输入序列的 embedding 的维度相同。

第三步：Norm 层到 Multi-head Attention 再过 Norm 层，通过多头子注意力模型。

第四步：MLP 将维度放大再缩小回去。

3.2 实验结果

在图像分类任务中，最为经典的是 CNN 模型[7]，这里着重把基于 Transformer 的 ViT 模型结果和 CNN 结果进行对比。

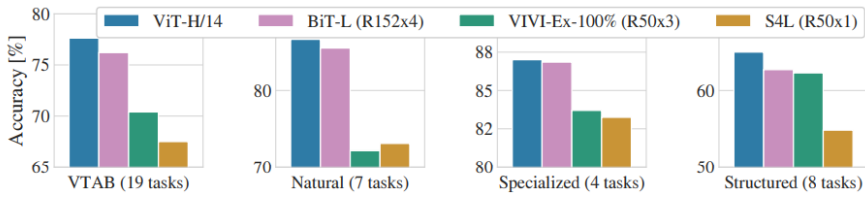

图 9　不同模型在 VTAB 上表现

图 9 实验结果表明在很大的数据集（VTAB）上预训练的时候，ViT 性能超越 CNN。

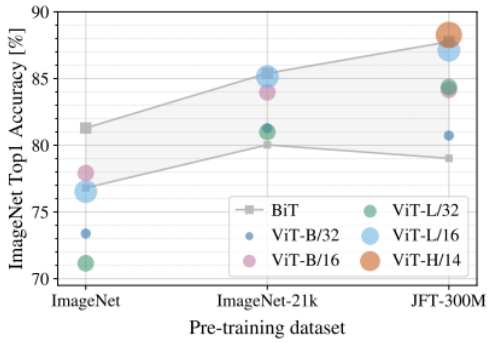

图 9　不同模型在 ImageNet 上表现

图 9 展示了在更小的数据集（ImageNet）上预训练的时候，ViT 微调的效果低于 ResNet，也就是说不是所有规模的数据集都适用于 ViT。

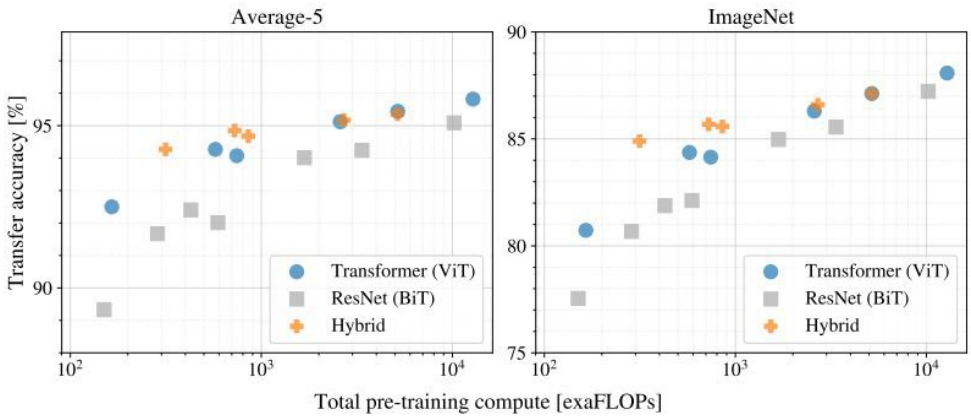

图 10　不同模型的表现和训练消耗对比

图 10 的结果证明了 ViT 的预训练比 ResNet 要消耗更小，相同训练复杂度的情况下，ViT 的效果可以好于 ResNet。

3.3 结果可视化

对于 Transformer 在图像视觉领域的应用，最吸引人的是 Transformer 模型是否真的像人类的视觉注意力机制那样形成一个注意力聚焦的过程，ViT 也对这个过程进行了可视化分析。

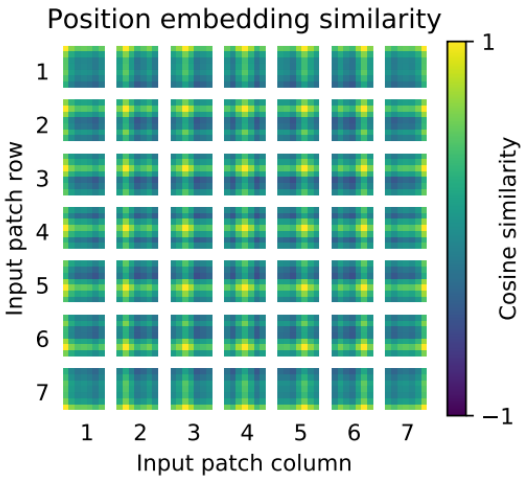

图 11 ViT 位置编码相似性分析

图 11 展示了 ViT 位置编码相似性的分析，位置越近，patches 之间的相似度越高，可以观察到相同行或者列的 patches 有着较为相似的 embeddings。

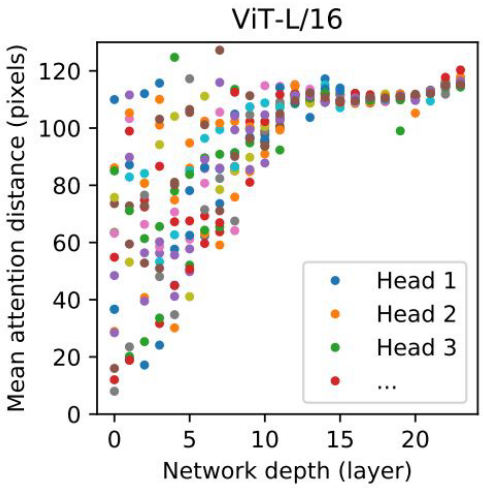

图 12 ViT 平均注意力距离

图 12 展示了 ViT 平均注意力距离随着网络深度变化的散点图，其中平均注意力距离是用算出的 attention weight 以及 query index 和其他所有的 pixel 求平均，来表示类似感受事业的概念，可以发现一些 head 在第一个 layer 就能够把 attention 注意到整张图片的范围。

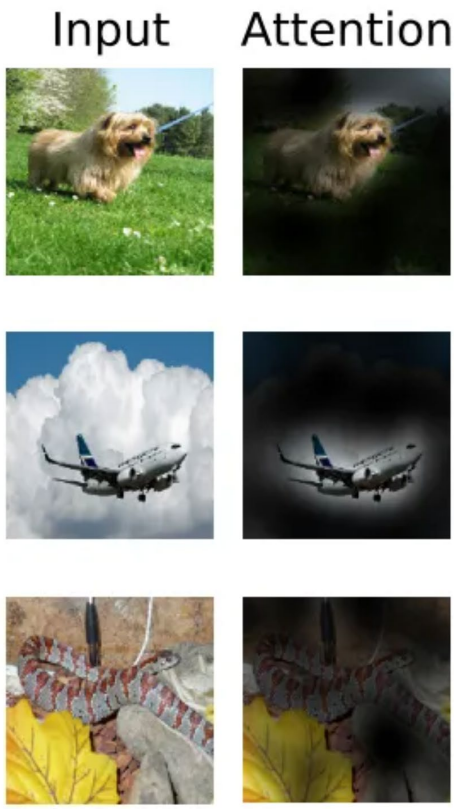

图 13 代表性的输入图像到输出的 token 里面的 attention 视野

图 13 通过可视化的方式形象展现了 attention 视野是如何对图像分类起到的作用，确实如视觉注意力机制所理解的那样，能够聚焦到图像中较为重要的特征部分进行分类分析。

3.4 小结

ViT 作为 Transformer 应用到二维图像领域里程碑的研究，提出了二维图像分类中，二维图像 2D 数据转换为 1D 数据输入 Transformer 模型的 Encoder 层的处理方法[8]。虽然小规模数据集下和 CNN 有一定的差距，但是也证明了 Transformer 在计算视觉领域的可行性，并且在模型训练消耗上和大数据集的应用上有着独有的优势。

4 Transformer 在三维视觉的应用-以 GeoTransformer 为例

以点云数据为代表的三维视觉近几年在自动驾驶等领域得到了充分的应用，尤其是在构建高精度地图时候需要对不同帧的点云数据进行配准，配准的目的是识别一个严格的转换函数实现对齐，建立两个点云之间的关联性，再利用关联性建立转换关系，使得不同帧的点云数据拼接起来，构成更为庞大的点云数据集合[9]。

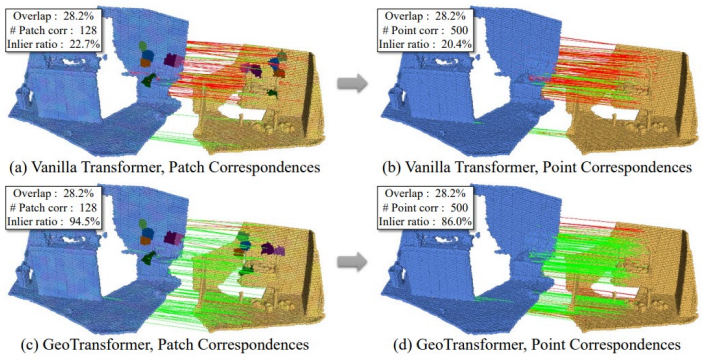

图 14 GeoTransformer 点云配准示意图

GeoTransformer 提出了一种基于 Transformer 引入全局结构信息[10]，实现了不需要 RANSAC 的场景点云配准，通过编码成对的距离和三重角，使得在点云数据低重叠的情况下也有较好的鲁棒性，并且具备刚性的变化不变性，能够实现极好的匹配精度，如图 14 所示。

4.1 GeoTransformer模型结构

整体来说 GeoTransformer 分为四个部分如图 15 所示：

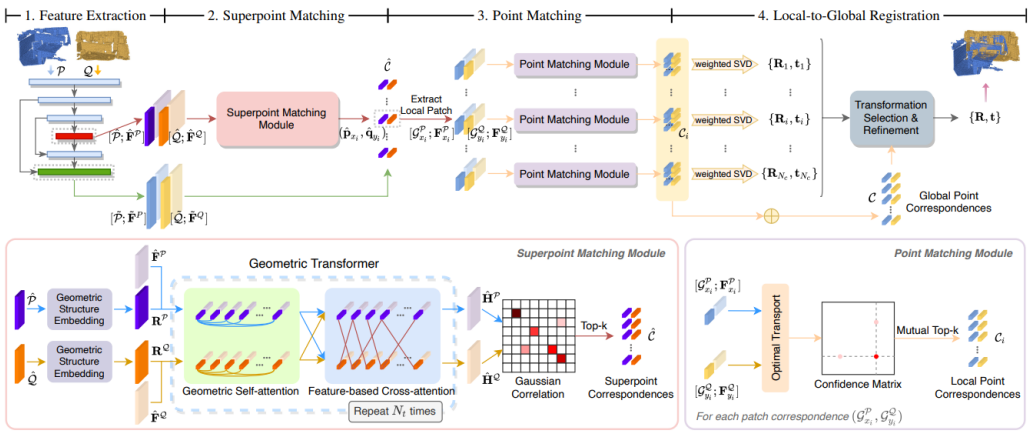

图 15 GeoTransformer 模型结构

第一部分：通过 KPConv backbone 从输入点云中下采样提取特征点[11]。

第二部分：通过建立的 superpoint matching 模块提取超点之间的关联属性。

第三部分：使用 point matching 模块对超点中对应的 patch 点进行匹配。

第四部分：使用 local-to-local 配准方法计算最终的配准转换函数。

其中作者主要提出了个改进的 geometric 自注意力模块如图 16 所示。

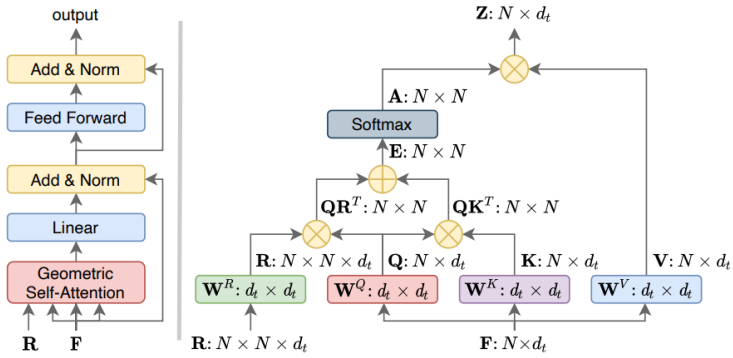

图 16 geometric 自注意力模块

现有的注意力结构没有很好的利用集合特征，只是聚合了高层的特征，特别是点云数据结构特殊，作者提出的方法是在自注意力的阶段，没有预先使用位置编码，而是 Q 和 K 计算的时候，对于 K 加入了几何的 Encoder 编码信息：

$$e_{i,j} = \frac{q_i(k_j + r_{i,j})}{\sqrt{d}} \tag{6}$$

这种几何信息的设计，满足了刚体结构的不变性，要求两个点云的特征提取满足共同的集合约束。

4.2 实验结果

本文主要在 3DMatch、3DLoMatch 和 KITTI 三个数据集上进行实验，可以发现 GeoTransformer 在 inlier ratio 这个指标上的提升，高质量的 correspondence 保证了很好的精度。

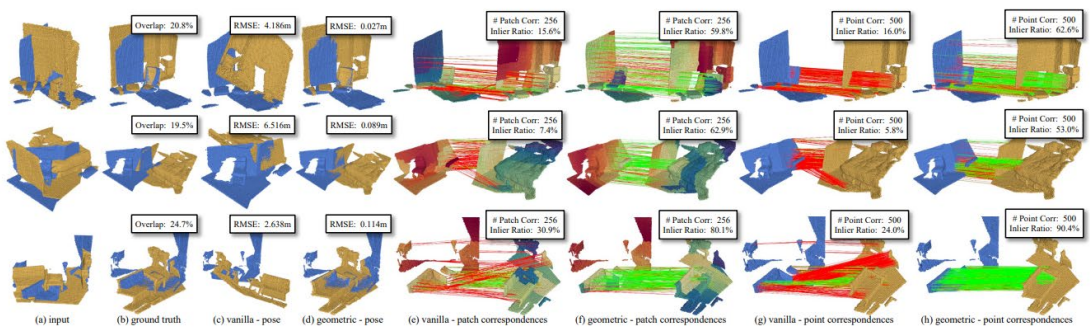

图 17 不同模型配准结果对比

图 17 也可视化展现出了即使在 overlap 很小的状态下，Geotransformer 仍然能够学习到非常一致的 attentionness。

5 总 结

Transformer 作为一种初始用于 NLP 领域的模型，在对 Encoder 编码模块或者对自注意力模块进行针对性的改进，可以很好的适应计算视觉中二维甚至三维数据特征的学习、提取和分析，针对于图像分类、三维点云配准多种任务中都能发挥较好的效果[12]。

References:

[1] VASWANI A, SHAZEER N, PARMAR N, 等. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[J]. Advances in neural information processing systems, 2014, 27.

[3] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.

[4] MT H, 戴葵. 神经网络设计[Z]. 北京: 机械工业出版社, 2002.

[5] HE K, ZHANG X, REN S, 等. Deep residual learning[J]. Image Recognition, 2015, 7.

[6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, 等. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[7] KOLESNIKOV A, BEYER L, ZHAI X, 等. Big Transfer (BiT): General Visual Representation Learning[C/OL]//VEDALDI A, BISCHOF H, BROX T, 等. Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 491-507. DOI:10.1007/978-3-030-58558-7_29.

[8] HAN K, WANG Y, CHEN H, 等. A Survey on Vision Transformer[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 1-1. DOI:10.1109/TPAMI.2022.3152247.

[9] KIM P, CHEN J, CHO Y K. SLAM-driven robotic mapping and registration of 3D point clouds[J]. Automation in Construction, 2018, 89: 38-48.

[10] QIN Z, YU H, WANG C, 等. Geometric transformer for fast and robust point cloud registration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11143-11152.

[11] THOMAS H, QI C R, DESCHAUD J E, 等. Kpconv: Flexible and deformable convolution for point clouds[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6411-6420.

[12] KHAN S, NASEER M, HAYAT M, 等. Transformers in vision: A survey[J]. ACM computing surveys (CSUR), 2022, 54(10s): 1-41.