ST-GCN模型说明

一、模型概述

空间 - 时间图卷积网络（Spatial Temporal Graph Convolutional Networks，ST-GCN）是一种用于基于骨骼的动作识别的创新模型。它将图神经网络扩展到时空图模型，通过自动学习数据中的空间和时间模式，有效克服了传统方法在建模骨骼时表达能力有限和难以泛化的问题。该模型在两个大型数据集Kinetics和NTU-RGB+D上展现出卓越性能，显著优于主流方法。

二、模型架构

（一）数据预处理与时空图构建

1. **数据来源与表示**
   - 基于骨骼的数据可从运动捕捉设备或视频姿态估计算法获取，通常为帧序列形式，每帧包含人体关节的2D或3D坐标。
   - 以Kinetics数据集为例，使用RTMposs工具箱估计视频中18个关节的2D坐标及置信度，将每个关节表示为(X, Y, C)的元组，一帧的骨骼数据即为18个这样元组的数组，多帧数据表示为$(17, 2, T)$维度的张量（T为帧数）。
2. **时空图构建**
   - 构建无向时空图$G = (V, E)$，其中节点集$V = \{v_{ti} | t = 1, \ldots, T, i = 1, \ldots, N\}$包含所有关节（$N$为关节数，$T$为帧数），节点特征向量$F(v_{ti})$由关节坐标向量和估计置信度组成。
   - 边集E由两部分组成：帧内基于人体结构自然连接的边$E_S = \{v_{ti}v_{tj} | (i, j) \in H\}$（$H$为自然连接的人体关节集）；连接连续帧相同关节的时间边$E_F = \{v_{ti}v_{(t+1)i}\}$。

（二）时空图卷积操作

1. **单帧内图卷积（空间图卷积）**
   - 借鉴2D图像卷积概念，将卷积操作扩展到图上。在单帧内，有N个关节节点V_t和相应的骨骼边$E_S(\tau)$。
   - 定义采样函数$p : B(v_{ti}) \to V$（$B(v_{ti})$为节点v_{ti}的邻居集，本文中取1-邻居集，即$d(v_{tj}, v_{ti}) \leq 1$的节点），在图上采样函数为$p(v_{ti}, v_{tj}) = v_{tj}$。
   - 权重函数$w(v_{ti}, v_{tj}) : B(v_{ti}) \to \mathbb{R}^c$通过对邻居集分区并映射标签实现。分区策略包括：
     - **单标签（Uni-labeling）**：最简单的策略，将整个邻居集视为一个子集，即所有邻居节点使用相同权重向量，类似（Kipf和Welling，2017）中的传

播规则，但会丢失局部差异特性，在单帧情况下相当于计算权重向量与所有邻居节点平均特征向量的内积。
    - **距离分区（Distance partitioning）**：根据节点到根节点的距离划分邻居集，如本文中设置$D=1$时，邻居集分为距离为0（根节点本身）和距离为1的两个子集，分别使用不同权重向量，能够建模局部差异属性，如关节间的相对平移。
    - **空间配置分区（Spatial configuration partitioning）**：依据节点与骨骼重心（一帧中所有关节平均坐标）的相对位置将邻居集分为三组：根节点本身、向心组（比根节点更靠近重心的邻居节点）、离心组（比根节点更远离重心的邻居节点），灵感来源于身体部位运动可分为同心和偏心运动，该策略在实验中表现较好。
  - 空间图卷积公式为$f_{out}(v_{ti})=\sum_{v_{tj}\in B(v_{ti})}\frac{1}{Z_{ti}(v_{tj})}f_{in}(p(v_{ti},v_{tj}))\cdot w(v_{ti},v_{tj})$，其中$Z_{ti}(v_{tj})$为归一化项，等于对应子集的基数，用于平衡不同子集对输出的贡献。
2. **时空建模（扩展到时空域）**
  - 定义时空邻居概念$B(v_{ti})=\{v_{qj}|d(v_{tj},v_{ti})\leq K,|q-t|\leq\lfloor\Gamma/2\rfloor\}$（$\Gamma$为时间核大小），将时间上连续的关节纳入邻居范围。
  - 采样函数与空间图卷积中相同，权重函数的标签映射修改为$l_{ST}(v_{qj})=l_{ti}(v_{tj})+(q-t+\lfloor\Gamma/2\rfloor)\times K$（$l_{ti}(v_{tj})$为单帧情况下的标签映射），从而实现时空图上的卷积操作。

（三）可学习边缘重要性加权

考虑到关节在不同身体部分运动中重要性不同，在每层时空图卷积添加可学习掩码M。掩码根据学习到的空间图边缘重要性权重，缩放节点特征对邻居节点的贡献。在实现中，将掩码与邻接矩阵进行元素相乘，如在单标签分区策略的图卷积公式中，将$A+I$替换为$(A+I)\otimes M$，掩码M初始化为全1矩阵，在训练过程中学习不同边缘的重要性权重。

（四）模型实现细节

1. **图卷积实现方式**
  - 采用与（Kipf和Welling，2017）类似的实现方式，用邻接矩阵A和单位矩阵I表示单帧内关节连接。对于单标签分区策略的ST-GCN，在单帧情况下可由公式$f_{out}=\Lambda^{-\frac{1}{2}}(A+I)\Lambda^{-\frac{1}{2}}f_{in}W$实现（其中$\Lambda^{ii}=\sum_j(A^{ij}+I^{ij})$，$W$为堆叠的权重矩阵）。
  - 在时空情况下，将输入特征图表示为(C,V,T)维度的张量，通过执行$1\times\Gamma$标准2D卷积并与归一化邻接矩阵$\Lambda^{-\frac{1}{2}}(A+I)\Lambda^{-\frac{1}{2}}$在第二维度相乘实现图卷积。
  - 对于多子集分区策略（如距离分区和空间配置分区），将邻接矩阵拆分为多个矩阵A_j（如距离分区中$A_0=I$，$A_1=A$），公式变为

$f_{out} = \sum_j \Lambda_j^{-\frac{1}{2}} A_j \Lambda_j^{-\frac{1}{2}} f_{in} W_j$（其中$\Lambda_j^{ii} = \sum_k (A_j^{ik}) + \alpha$，本文中$\alpha = 0.001$以避免$A_j$中出现空行）。

2. **网络架构与训练参数**
    - ST-GCN模型由9层时空图卷积算子（ST-GCN单元）组成，前三层输出64通道，中间三层输出128通道，最后三层输出256通道，时间核大小为9。
    - 每个ST-GCN单元应用Resnet机制，并在之后以0.5概率随机失活（dropout）特征以避免过拟合。第4层和第7层时间卷积层步长设为2作为池化层。
    - 模型使用随机梯度下降训练，学习率为0.01，每10个epoch衰减0.1。在Kinetics数据集训练时，为避免过拟合进行两种数据增强：随机仿射变换（模拟相机运动）和随机采样片段（训练时随机采样，测试时使用全部帧）。

三、模型优势

1. **强大的表达能力**
    - 通过自动学习空间和时间模式，能更有效地捕捉骨骼序列中的动态信息，相比传统依赖手工特征或规则的方法，能更好地建模动态骨骼，从而提升动作识别的准确性。
2. **良好的泛化能力**
    - 模型架构不依赖于特定的关节数量或连接方式，可在不同数据集上工作，如在Kinetics（2D关节坐标）和NTU-RGB+D（3D关节坐标）数据集上均取得优异性能，表明其能适应多种数据场景。
3. **实验性能卓越**
    - 在Kinetics和NTU-RGB+D两个大规模数据集上的实验表明，ST-GCN在动作识别准确率上相比之前的方法有显著提升，在NTU-RGB+D数据集的跨主体（X-Sub）和跨视角（X-View）基准测试中均达到领先水平。

四、模型应用示例

1. **在Kinetics数据集上的应用**
    - 输入视频经姿态估计后构建时空图，通过ST-GCN模型的多层时空图卷积操作生成高级特征图，最后由SoftMax分类器预测动作类别。在Kinetics数据集的实验中，通过与多种方法对比，ST-GCN在top - 1和top - 5分类准确率上表现出色，验证了其在大规模无约束动作识别任务中的有效性。
2. **在NTU-RGB+D数据集上的应用**
    - 该数据集为约束环境下采集，具有固定的相机视角和3D关节标注。ST-GCN模型在该数据集上同样表现优异，在不使用数据增强的情况下，超越了之前的多种方法，进一步证明了模型的泛化能力和有效性。同时，通过对Kinetics和NTU-RGB+D数据集的实验对比，展示了ST-GCN在不同数据特性下的适应性和优势。

五、总结

ST-GCN模型为基于骨骼的动作识别提供了一种有效且通用的解决方案。其创新的时空图卷积架构、自动学习模式的能力以及在不同数据集上的卓越表现，使其在动作识别领域具有重要意义。未来可进一步探索如何将上下文信息（如场景、对象和交互）融入模型，以进一步提升性能。