ST-GCN 时空图卷积网络模型

作者：颜思捷，熊元骏，林达华

文章链接：

Github 代码：

简介

近日，香港中大-商汤科技联合实验室的最新 AAAI 会议论文「Spatial Temporal Graph Convolution Networks for Skeleton Based Action Recognition」提出了一种新的 ST-GCN，即时空图卷积网络模型，用于解决基于人体骨架关键点的人类动作识别问题。该方法除了思路新颖之外，在标准的动作识别数据集上也取得了较大的性能提升。本文中我们将详细介绍该论文中提出的方法，并介绍一些计划中的进一步工作等。

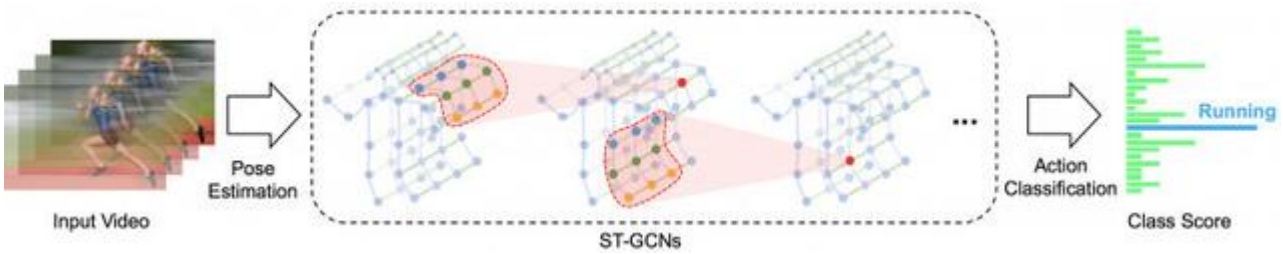

图 1 ST-GCN 的模型结构示意图

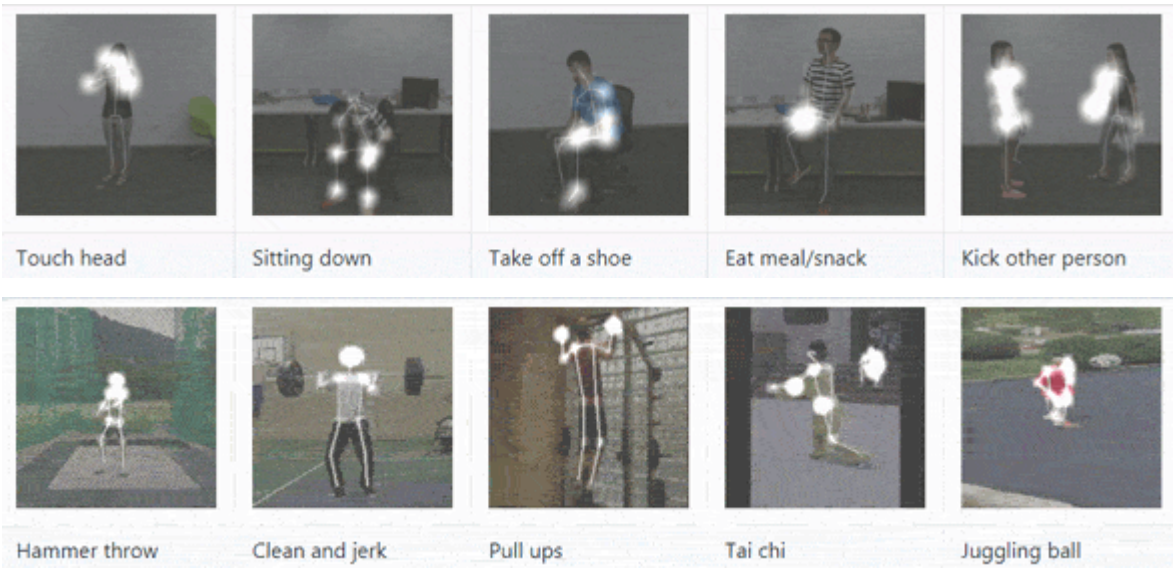

图 2 对 ST-GCN 最末卷积层的响应进行可视化的结果。（via GitHub）

基于骨架关键点的动作识别

随着如 Microsoft Kinect、OpenPose 等人体姿态检测系统的成熟，基于骨架关键点的人类动作识别成了计算机视觉，特别是人类动作识别研究中的一个重要任务。该任务要求输入在连续的视频帧中检测到的人体骨架关键点序列，输出正在发生的人类动作类别。作为动作识别系统中的重要模态，基于骨架的动作识别已经展现出重要的实用价值与研究价值。本论文正是针对这个任务提出了一种全新的深度学习模型，我们称之为「时空图卷积网络」（ST-GCN）。

构建时空图

ST-GCN 的基础是时空图结构。从骨架关键点序列构建时空图 (spatial-temporal graph) 的想法来源于我们对现有的骨架动作识别方法以及图像识别方法的观察。我们发现，现有的基于骨架的动作识别方法中为了提高识别精度多数引入了一些空间结构信息，包括相邻关键点的连接关系或身体部件等（如手-手肘-肩膀的连接关系）。

为了建模这些空间信息，现有方法常常使用 RNN 等序列模型来遍历相连的关键点。这就要求模型设计者定义一种遍历的规则，或者手动定义一些身体部件。我们指出，在这种设计中，很难得到一个最优的遍历规则或者部件划分。但是，我们发现，关键点之间天然的连接关系，其实构成了一个天然的图结构（graph）。那么，我们怎么能够高效地使用这些图结构来进行动作识别呢?

在 ST-GCN 的工作中我们提出，可以从输入的关键点序列中建立一个时空图（spatial-temporal graph）。这个图结构按照如下的规则来构建。

\1. 在每一帧内部，按照人体的自然骨架连接关系构造空间图；2. 在相邻两帧的相同关键点连接起来，构成时序边；3. 所有输入帧中关键点构成节点集（node set），步骤 1、2 中的所有边构成边集（edge set），即构成所需的时空图。

在按照上述规则得到的时空图上，我们自然地保留了骨架关键点的空间信息，并使得关键点的运动轨迹（trajectory）以时序边的形式得到表现。这使得我们可以设计一个统一的模型来完整地对这些信息进行建模。在图 3 中我们展示了一种时空图的结构。

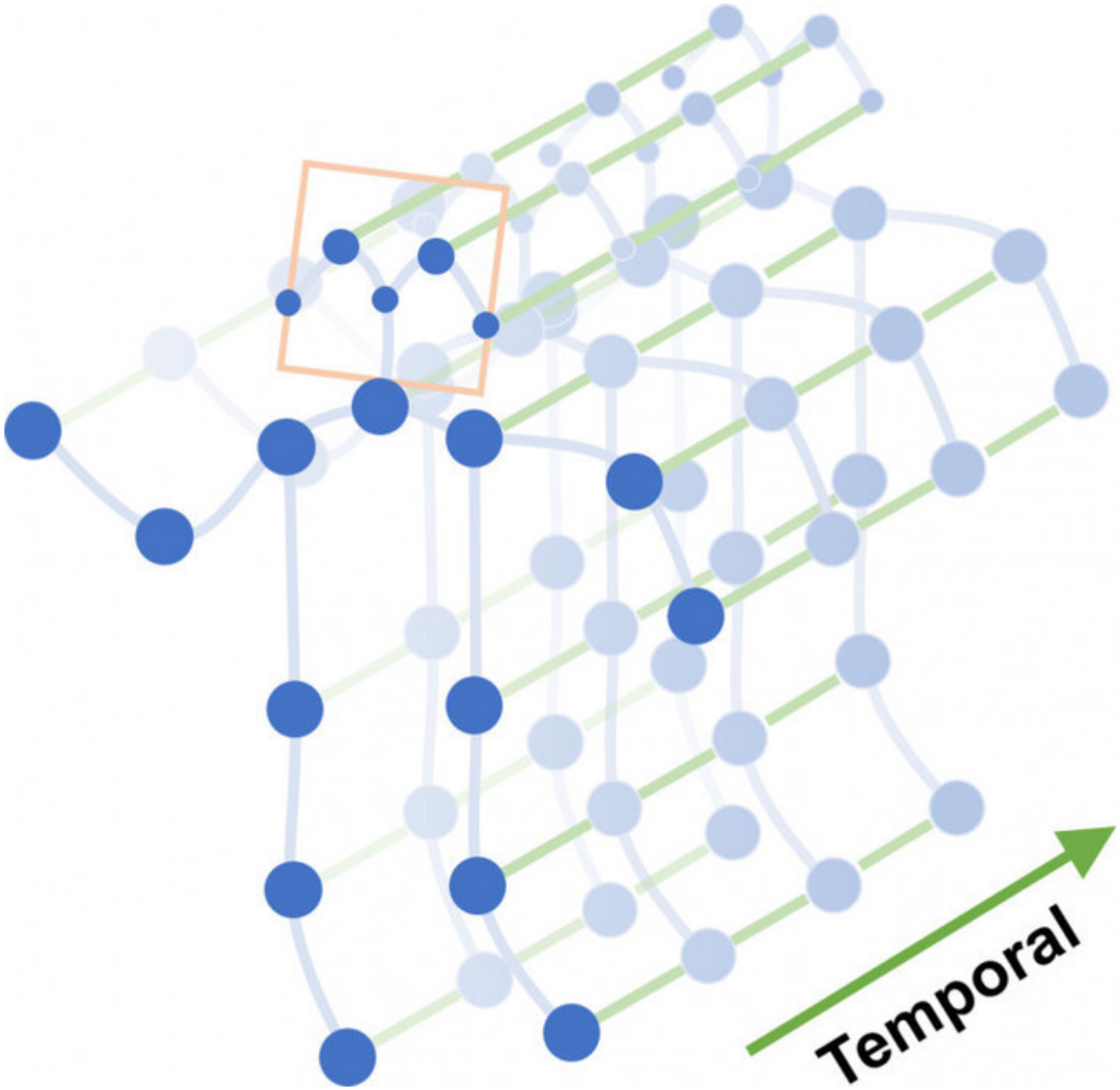

图 3. NTU-RGBD 数据集上建立的时空图示例。

图结构上的卷积网络

为了在时空图上对人类动作的信息进行分析，我们提出使用图卷积网络 (graph convolutional networks – GCN)。图上的神经网络模型是机器学习研究的一个热点领域。本文中使用的图卷积网络即是图上神经网络中的一种，其在网络分析、文本分类等问题都有成功应用。

在介绍图卷积网络的概念之前，我们先来回顾图像上的卷积操作。在图像上，卷积操作使用一些固定大小的卷积核（filter/kernel）来扫描输入的图像。如图 3 所示，在每个扫描的中心位置像素附近，抽取一个与权重矩阵大小相同的像素矩阵，将这些像素上的特征向量按空间顺序拼接并与卷积核的参数向量做内积以得到该位置的卷积输出值。在这里，「附近像素」可以定义为像素网格（grid）上的一个邻域（neighborhood）。将图像上的卷积操作推广到任意结构的图结构上时，我们同样可以定义任何一个节点的邻域，与一系列权重矩阵。这就是图卷积网络的基本思想。

但是，与图像不同的是，普通的图结构上如果使用邻接矩阵（Adjacency matrix）来定义邻域时，每个节点的邻域中节点的数量并不是固定的（考虑补 0 时，图像上像素附近的像素是总是固定的）。这就使得我们我们很难确定：1）需要使用的卷积核的参数维度；2）如果排列权重矩阵与邻域内的节点以进行内积运算。在原始的 GCN 文章中，作者提出了将内积操作变为这样一个操作：使用同一个向量与所有邻域内的点上的特征向量计算内积并将结果求均值。这使得： 1）卷积核的参数可以确定为一个固定长度的向量；2）不需要考虑邻域内节点的顺序。这个设计使得 GCN 可以在任意连接关系的图上使用，并在一些任务，如网络分析和半监督学习中取得了较好的性能。

需要注意的是，图上神经网络除了上面提到基于图的空间结构的思路之外，还有一种基于谱分析 (spectral analysis) 的构造思路。关于这一类方法，请见参考文献【2】。在 ST-GCN 中，我们也使用了基于图的空间结构的思路。

时空图卷积网络与动作识别

要将图卷积网络运用于基于骨架关键点的动作识别中，我们还需要仔细分析这个任务的特点与难点，而不是直接将已有方法生搬硬凑。在本文中，我们指出了原始 GCN 的一个重要性质：该文中提出的卷积操作，实质上等价于先将邻域内所有节点的特征向量求平均，再与卷积核的参数向量计算内积。这种平均操作在骨架动作识别会遇到一个重要问题，即：它无法建模关键点之间相对位置变化的情况，或所谓的「微分性质」（differential properties）。因此基于原始 GCN 的模型，识别性能并不会很理想。

针对这个问题，我们认为，要真正增强的模型的性能，必须跳出原始 GCN 的「平均思想」。为了解决这个问题，我们将理解了图像上的卷积操作理解为把中心像素相邻的像素集合（邻域集-neighbor set）按照空间顺序，如从左至右，从上至下，划分为一系列集合。在图像的情形中，每个集合正好包含一个像素。这些集合就构成了邻域集的一个划分（partition）。卷积核的参数只与这个划分中的子集个数以及特征向量长度有关。那么在普通的图结构中，只要定义了某种划分规则（partitioning strategy），我们就也可以参照图像卷积来定义卷积核的参数。类似的思想也应用在了如 deformable CNN 等近期工作中。

有了这个思想，我们就可以针对骨架动作识别，乃至任何图卷积网络所面对的问题来定义有针对性的卷积操作。而定义卷积操作就简化为了设计对应的划分规则。对一个存在 K 个子集的划分规则，卷积核的参数包含 K 个部分，每个部分参数数量与特征向量一样。仍然以图像上的卷积为例，在一个窗口大小为 3x3 的卷积操作中，一个像素的邻域（窗口）按照空间顺序被划分为 9 个子集（左上，上，右上，左，中，右，左下，下，右下），每个子集包含一个像素。卷积核的参数包含 9 个部分，每个部分与特征图（feature map）的特征向量长度（number of channel）一致。也就是说，图像卷积可以解释为普通图上卷积在规则网格图（regular grid）上的一种应用。

为了在时空图上进行骨架动作识别，我们提出了三种空间的划分规则。

第一种称为「唯一划分」(uni-labeling)。其与原始 GCN 相同，将节点的 1 邻域划分为一个子集。

第二种称为「基于距离的划分」(distance partitioning)，它将节点的 1 邻域分为两个子集，即节点本身子集与邻节点子集。引入基于距离的划分使得我们可以分析骨架关键点之间的微分性质。

进一步，我们针对动作识别的特点，提出了第三种，「空间构型划分」(spatial configuration partitioning)。这种划分规则将节点的 1 邻域划分为 3 个子集，第一个子集为节点本身，第二个为空间位置上比本节点更靠近整个骨架重心的邻节点集合，第三个则为更远离重心的邻节点集合。建立这种划分规则在根据运动分析的研究中对向心运动与离心运动的定义。三种划分规则的示意图请见图 4。

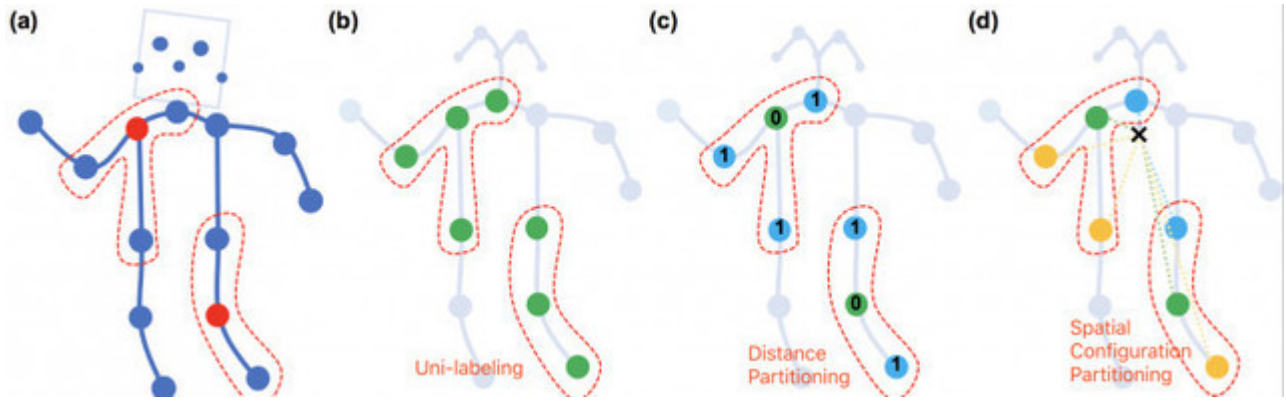

图 4，三种空间的划分规则示意图。

除了同一帧内部的空间划分规则，在时间上，由于时序边构成了一个网格，我们可以直接使用类似于时序卷积（temporal convolution）的划分规则。最终，时空图上使用的划分规则得到的子集集合会是空间划分与时序划分的笛卡尔积。

定义好了时空图上的卷积操作，我们就可以设计卷积网络了。为了展示 ST-GCN 的性能，我们直接从一个已有的时序卷积网络结构的基础上设计了文中用到的 ST-GCN 的网络结构。我们将所有时序卷积操作转为时空图的卷积操作，每一个卷积层的输出是一个时空图，图上每一个节点保有一个特征向量。最终，我们合并所有节点上的特征并使用线性分类层进行动作分类。训练使用标准的 SoftMax 交叉熵损失函数进行监督。参数学习使用标准随机梯度下降算法（SGD）。

实验结果

我们在两个性质迥异的骨架动作识别数据集上进行了实验来验证 ST-GCN 的性能。

第一个数据集是 Kinetics-Skeleton，它来自于最近由 Google DeepMind 贡献的 Kinetics 视频人类动作识别数据集。我们使用 OpenPose 姿态估计软件得到视频中所有的骨架关键点信息来构成 Kinetics-Skeleton。该数据集共有约 30 万个视频与 400 类动作。

第二个数据集是 NTU-RGB+D，这是三维骨架动作识别的标准测试数据集。它包含了用 Microsoft Kinect 采集的三维骨架序列。该数据集共有约 6 万个视频，60 个动作类别。这个数据集包含了两个测试协议，即跨表演人（X-Sub）与跨视角（X-View）协议。在两个数据集的三个测试协议上，ST-GCN 相比现有方法在识别精度上均有很大提高，具体结果可见表图 1。

	Top-1	Top-5
RGB(Kay et al. 2017)	57.0%	77.3%
Optical Flow (Kay et al. 2017)	49.5%	71.9%
Feature Enc. (Fernando et al. 2015)	14.9%	25.8%
Deep LSTM (Shahroudy et al. 2016)	16.4%	35.3%
Temporal Conv. (Kim and Reiter 2017)	20.3%	40.0%
ST-GCN	**30.7%**	**52.8%**

Table 2: Action recognition performance for skeleton based models on the Kinetics dataset. On top of the table we list the performance of frame based methods.

	X-Sub	X-View
Lie Group (Veeriah, Zhuang, and Qi 2015)	50.1%	52.8%
H-RNN (Du, Wang, and Wang 2015)	59.1%	64.0%
Deep LSTM (Shahroudy et al. 2016)	60.7%	67.3%
PA-LSTM (Shahroudy et al. 2016)	62.9%	70.3%
ST-LSTM+TS (Liu et al. 2016)	69.2%	77.7%
Temporal Conv (Kim and Reiter 2017).	74.3%	83.1%
C-CNN + MTLN (Ke et al. 2017)	79.6%	84.8%
ST-GCN	**81.5%**	**88.3%**

Table 3: Skeleton based action recognition performance on NTU-RGB+D datasets. We report the accuracies on both the cross-subject (X-Sub) and cross-view (X-View) benchmarks.

表图 1. 骨架动作识别结果

除了得到更好的性能，我们也详细分析了三种划分规则对识别精度的影响。如表 2 所示，正如我们所期望的，距离划分与空间构型划分相对于原始 GCN 使用的唯一划分在精度上均有较大提高。这证明了引入新的划分规则的重要性。特别的，针对动作识别任务设计的空间构型划分取得了最高的性能，并被最后应用于 ST-GCN 的相关实验中。

我们还将 ST-GCN 的最后一层神经元响应进行了可视化（表 2）。在结果中我们可以明显看到 ST-GCN 能够追踪并深入分析在某个时间段与动作最相关的身体部分的运动，这解释了为何 ST-GCN 相对于其他不关注空间结构的现有方法能得到很大的性能提高。

	Top-1	Top-5
Baseline TCN	20.3%	40.0%
Local Convolution	22.0%	43.2%
Uni-labeling	19.3%	37.4%
Distance partitioning*	23.9%	44.9%
Distance Partitioning	29.1%	51.3%
Spatial Configuration	29.9%	52.2%
ST-GCN + Imp.	**30.7%**	**52.8%**

Table 1: Ablation study on the Kinetics dataset. The "ST-GCN+Imp." is used in comparison with other state-of-the-art methods. For meaning of each setting please refer to Sec.4.2.

表图 2，不同划分规则的影响

思考与最后的话

回顾 ST-GCN 的提出，我们总结了两个重要的思想跨越。

第一个是从将骨架序列理解为一帧帧的骨架演进为将整个视频理解为一个整体的时空图，这使得用一个统一的模型来分析动作成为可能。第二个是从原始 GCN 的朴素思想演进为使用基于划分规则的卷积定义。这个思想使得我们可以超越原始 GCN 并得到巨大的性能提升，该思想也在 MoNet【3】的工作中被提及过。我们将其原则化为 集合的划分操作。这也使得这个思想可以应用其他的分析任务中。

在将来的工作中，我们计划运用 ST-GCN 的灵活性来处理更多的图分析问题。同时，针对动作识别任务，一个自然的演进就是在骨架关键点坐标的基础上引入视觉特征，如图像特征，乃至场景图（scene-graph）等，并将它们统一在 ST-GCN 的分析框架下。我们的最终目标则是一个性能更高，更具有可解释性的统一的视频动作识别模型。

相关文献：

【1】「Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition」, Sijie Yan, Yuanjun Xiong and Dahua Lin, AAAI 2018.

【2】「Convolutional neural networks on graphs with fast localized spectral filtering.」， Defferrard, et. al., NIPS 2016.

【3】 "Geometric deep learning on graphs and manifolds using mixture model CNNs.", Monti, Federico, et al. CVPR 2017.