Two-stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition （2S-AGCN）论文简略笔记

该文提出的网络中文可以翻译为 双流自适应图卷积网络(Two-stream Adaptive Graph Convolutional Networks)

关键词：two-stream， adaptive， Graph Convolutional Network

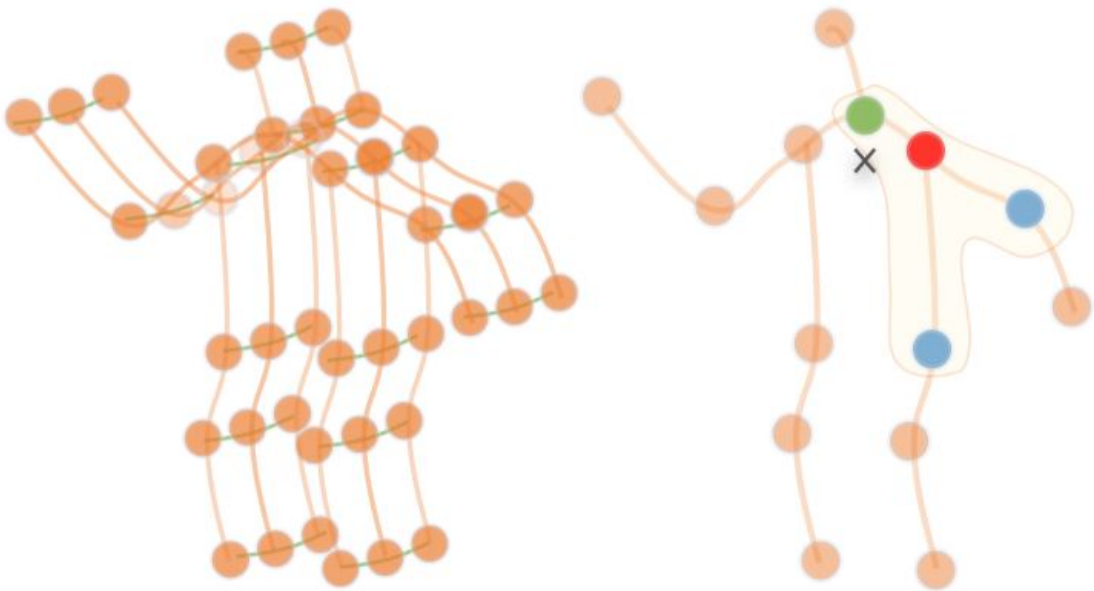
图卷积网络(Graph Convolutional Network): 该图卷积网络基本架构来源于15年的ST-GCN。

自适应(adaptive):ST-GCN的边连接是基于人体骨骼结构设计的，是固定的。而这种手工设计在分类任务中未必是最优的，因此作者额外加入权重矩阵进行训练。

双流(two-stream)：除了一阶信息(关节点joint坐标)外，二阶信息(骨头bone的坐标以及方向）对行为识别任务也能提供许多信息，因此作者对于搭建的网络会分别放入两类输入训练，最终将两个模型融合(fusion)。

图卷积网络(Graph Convolutional Network)

骨架图如下所示，与ST-GCN相同：将人体重心点作为根节点，每个节点都有邻接点，根据跟根节点的距离分为三类——自己本身，相对自己离根节点更近，相对自己离根节点更远。



图卷积形式如下: f为特征图，v为节点，B为邻接节点集合，w为权重，l代表节点特有的权重，Z是集合B节点数目。

$$f_{out}(v_i) = \sum_{v_j \in \mathcal{B}_i} \frac{1}{Z_{ij}} f_{in}(v_j) \cdot w(l_i(v_j))$$

具体实现如下：

K指邻接节点分为几类，如前文所言本文根据与根节点的距离设为3类，

A可以视作是邻接矩阵，W是权重矩阵，M是attention map

$$\mathbf{f}_{out} = \sum_{k}^{K_v} \mathbf{W}_k (\mathbf{f}_{in} \mathbf{A}_k) \odot \mathbf{M}_k$$

自适应(Adaptive)

自适应的目的是为了使得图结构不是固定的，能够根据数据训练出一个比手工权重更适合分类的结构。

A与原本的A一致

B为由数据训练出来的权重矩阵，采用加法是为了能够让原本为零的权重不为零。因为如果图不存在边，A本身是零，无论B的值为多少，AB相乘也必然是零。

C用于计算两个节点特征的相似度，范围为0-1

$$\mathbf{f}_{out} = \sum_{k}^{K_v} \mathbf{W}_k \mathbf{f}_{in} (\mathbf{A}_k + \mathbf{B}_k + \mathbf{C}_k)$$

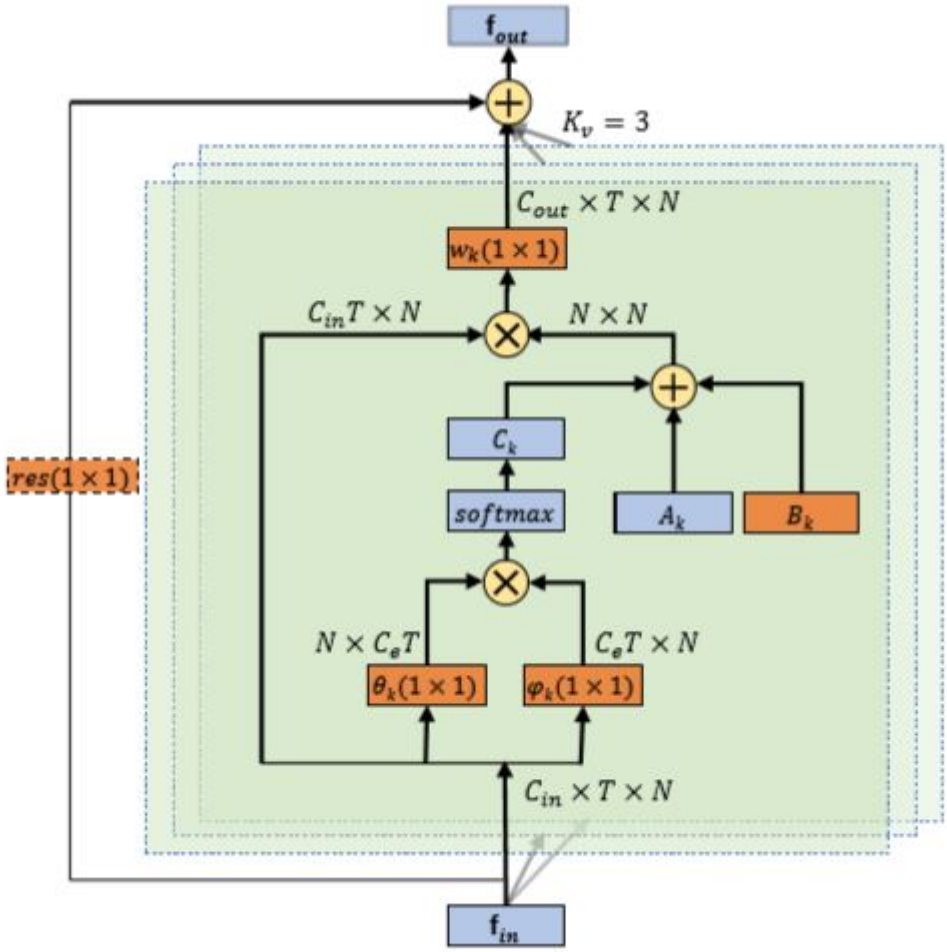
C的值由归一化高斯函数计算所得，如下：

$$f(v_i, v_j) = \frac{e^{\theta(v_i)^T \phi(v_j)}}{\sum_{j=1}^{N} e^{\theta(v_i)^T \phi(v_j)}}$$

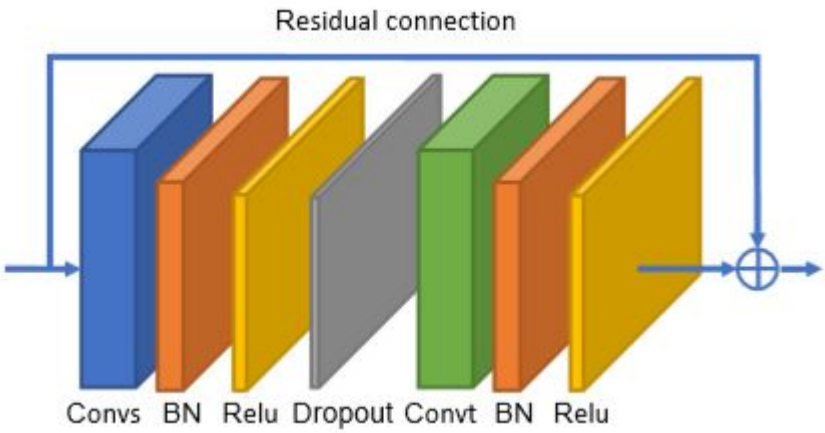
具体实现如下：归一化操作即softmax

$$\mathbf{C}_k = softmax(\mathbf{f_{in}}^T \mathbf{W}_{\theta k}^T \mathbf{W}_{\phi k} \mathbf{f_{in}})$$

因此最终adaptive GCN layer结构如下所示



block 如下所示：



双流(two-stream)

除了一阶信息(关节点joint坐标)外，二阶信息(骨头bone的坐标以及方向）对行为识别任务也能提供许多信息，因此作者对于搭建的网络会分别放入两类输入训练，最终将两个模型融合(fusion)。

对于二阶信息，骨头的坐标为骨头的两个点坐标相减。靠近根节点的为源节点，远离根节点的为目标节点，后者减去前者即为骨头坐标的信息。

因为骨架图是非循环图/树，所以最终骨头的数目会比关节少一个，为了简化网络设计补一个(0,0,0)。

实验结果

消融实验：证明自适应模块的作用：

Methods	Accuracy (%)
ST-GCN	92.7
ST-GCN wo/M	91.1
AGCN wo/A	93.4
AGCN wo/B	93.3
AGCN wo/C	93.4
AGCN	93.7

证明双流确实有效

Methods	Accuracy (%)
Js-AGCN	93.7
Bs-AGCN	93.2
2s-AGCN	95.1

跟其它SOTA方法在NTU 和Kinetics数据集上比较

Methods	X-Sub (%)	X-View (%)
Lie Group [31]	50.1	82.8
HBRNN [6]	59.1	64.0
Deep LSTM [27]	60.7	67.3
ST-LSTM [22]	69.2	77.7
STA-LSTM [29]	73.4	81.2
VA-LSTM [33]	79.2	87.7
ARRN-LSTM [19]	80.7	88.8
Ind-RNN [20]	81.8	88.0
Two-Stream 3DCNN [21]	66.8	72.6
TCN [14]	74.3	83.1
Clips+CNN+MTLN [13]	79.6	84.8
Synthesized CNN [23]	80.0	87.2
CNN+Motion+Trans [18]	83.2	89.3
3scale ResNet152 [17]	85.0	92.3
ST-GCN [32]	81.5	88.3
DPRL+GCNN [30]	83.5	89.8
2s-AGCN (ours)	88.5	95.1

Methods	Top-1 (%)	Top-5 (%)
Feature Enc. [8]	14.9	25.8
Deep LSTM [27]	16.4	35.3
TCN [14]	20.3	40.0
ST-GCN [32]	30.7	52.8
Js-AGCN (ours)	35.1	57.1
Bs-AGCN (ours)	33.3	55.7
2s-AGCN (ours)	36.1	58.7

小结

该论文最主要的两个点分别为自适应和双流，自适应使得能够自主的构建图结构，双流利用了骨头的信息。

发现这些想法其实有很多论文也有，所以读论文的时候自己要注意吸收然后尝试组合做实验。

比如这篇文章 Skeleton-Based Action Recognition with Directed Graph Neural Network(cvpr2018) 也使用了自适应模块和骨头的信息，并且在NTU-RGB的数据集上效果略好过本文(cvpr 2019)。