《Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting》

《时空图卷积网络：交通预测的深度学习框架》

摘要：

1. **本文的背景：**由于交通流的高度非线性和复杂性，传统方法，例如 k 最近邻算法（KNN）、支持向量机（SVM），无法满足中长期预测任务的要求，且往往忽略了数据的时空依赖性。因此，为了解决在交通领域的时间序列预测问题，本文试图提出一种新颖的深度学习框架。
2. **本文的贡献：**提出一种新型的深度学习框架——时空图卷积网络（STGCN），用于对空间和时间依赖关系进行建模。而这也是第一次在交通研究中应用纯卷积结构从图结构时间序列中同时提取时空特征。
3. **主要创新点：**设计了一种新型的时空卷积块（ST-Conv blocks），而每个块包含两个时间门控卷积层（Temporal Gated-Conv）和一个位于中间的空间图卷积层（Spatial Graph-Conv）。
4. **实验结果：**通过在两个真实世界的交通数据集上的实验结果表明，$STGCN$模型在具有多个预设预测长度和网络尺度的预测任务中始终优于现有的基准模型。

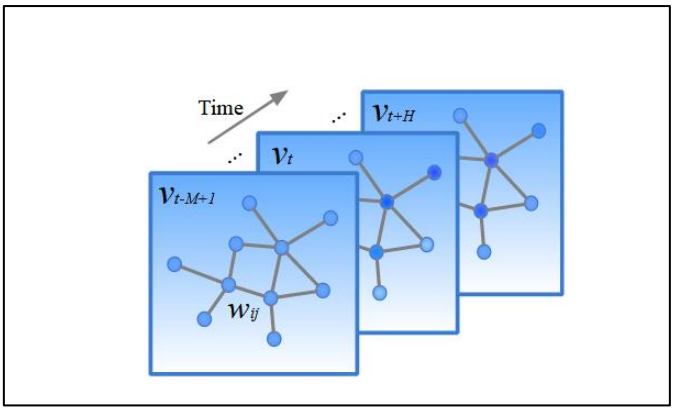
问题定义：

交通预测是一个典型的时间序列预测问题，即在给定前M个交通观测值的情况下，预测未来H个时间步长中最有可能的交通测量值（例如速度或交通流量），公式表达如下所示：

$$\hat{v}_{t+1}, \ldots, \hat{v}_{t+H} = \underset{v_{t+1},\ldots,v_{t+H}}{\arg\max} \log P(v_{t+1}, \ldots, v_{t+H} | v_{t-M+1}, \ldots, v_t), \quad (1)$$

其中，$v_t \in \mathbb{R}^n$是时间步长t处n个路段的观测向量，而每个元素都记录了单个路段的历史观测值。

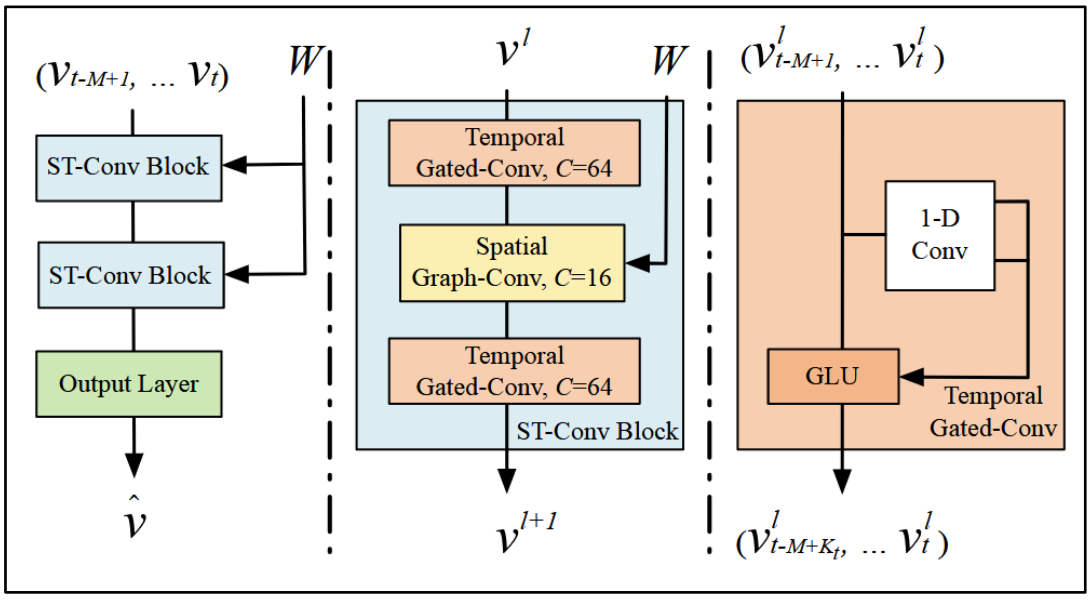
此外，我们还需要在图（Graph）上定义交通网络，因此各路段的观测值v_t不是独立的，而是通过图结构进行进一步表示的。具体情况如下图所示：



我们用G表示在本篇研究中所定义的时空图，而在第t个时间步长处的图定义为$G_t = (V_t, E, W)$。其中，V_t是一组有限的顶点集，它与交通网络中n个监测站的观测结果相对应；E为一组边集，表示站点之间的连通性；而$W \in \mathbb{R}^{n \times n}$表示$G_t$的加权邻接矩阵。**注意：不同时间步的$W$不变。**

$STGCN$模型框架：

$STGCN$网络整体架构如下图所示：



可以看到，$STGCN$由多个时空卷积块组成，每个时空卷积块又形成了一个“三明治”结构，两边各有一个时序门控卷积层，而中间夹着一个空间图卷积层。模型的输入结构是M个时间步的图的特征向量$X \in \mathbb{R}^{M \times n \times C_i}$（论文中$C_i = 1$）以及对应的邻接矩阵$W \in \mathbb{R}^{n \times n}$，然后在经过两个时空卷积块和一个输出层后，输出结果$\hat{v} \in \mathbb{R}^n$来预测第t个时间步后某个时间步特征。

各模块的详细内容如下：

1. 时域卷积块

时域卷积块如上图最右侧所示，它沿着时间维度进行一维卷积，卷积核大小为K_t，并且由门控线性单元（GLU）作为非线性激活函数。因此，对于图G中的每个节点，时间卷积块会在没有填充的情况下探索每个输入元素的K_t跳时域邻居。所以，这会导致每次序列长度缩短K_t-1。

简而言之，每个节点的时间卷积输入可以看作是一个长度为M、通道数为C_i的时间序列$Y \in \mathbb{R}^{M\times C_i}$。卷积核$\Gamma \in \mathbb{R}^{K_t\times C_i\times 2C_o}$，它将输入$Y$映射到单个输出元素，从而得到$[P\ Q] \in \mathbb{R}^{(M-K_t+1)\times 2C_o}$。（其中，P、Q 在相同大小的通道下被分成两半）。然后，再进行GLU激活。

因此，时间门控卷积可以被表示为：

$$\Gamma *_{\mathcal{T}} Y = P \odot \sigma(Q) \in \mathbb{R}^{(M-K_t+1)\times C_o},$$

而对于一张完整的时空图：**输入$\boldsymbol{X} \in \mathbb{R}^{M\times n\times C_i}$，输出$\boldsymbol{Y} \in \mathbb{R}^{(M-K_t+1)\times n\times C_o}$**。

2. 空域卷积块

简单来说，空域卷积是在每个时间步的图上进行$ChebGCN$（不在时间步之间进行）。因此，它的输入为$X \in \mathbb{R}^{n\times C_i}$，输出为$Y \in \mathbb{R}^{n\times C_o}$。而具体的实现公式如下：

$$Y \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{L})X$$

其中，$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$，$T_0(x) = 1$，$T_1(x) = x$；$\tilde{L} = \frac{2}{\lambda_{max}}L - I_N$，$L = I_N - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = U\Lambda U^T$，$\lambda_{max}$是$L$的最大特征值；$\theta \in \mathbb{R}^K$是切比雪夫多项式系数组成的向量。【$I_N$为单位矩阵，$D$、$A$和$\Lambda$分别为图的度矩阵、邻接矩阵和特征值的对角矩阵。】

而在本文中，$K=3$，卷积核$\Theta \in \mathbb{R}^{K\times C_i}$，个数为$C_o$。因此，对于一张完整的时空图：**输入$\boldsymbol{X} \in \mathbb{R}^{M\times n\times C_i}$，输出$\boldsymbol{Y} \in \mathbb{R}^{M\times n\times C_o}$**。

补充：

如果我们在空域卷积中直接采用GCN的方式，则上述相关公式可进一步简化为：

$$Y = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta)$$

其中，$X \in \mathbb{R}^{n \times C_i}$，$\Theta \in \mathbb{R}^{C_i \times C_o}$，$\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} \in \mathbb{R}^{n \times n}$，$\sigma$为激活函数，因此输出结果$Y \in \mathbb{R}^{n \times C_o}$。（**注意：**这里指的是每一个时间步下的空域卷积。）

3. 输出层

根据时域卷积块的一维卷积，每经过一个时空卷积块，数据在时间维度的长度就会减小$2(K_t - 1)$。因此，经过两个时空卷积块后，输出$\boldsymbol{Y \in \mathbb{R}^{(M-4(K_t-1)) \times n \times C_o}}$，而它又会作为输出层的输入。

$STGCN$的输出层包括一个时域卷积层和一个全连接层，时域卷积层的卷积核大小$\Gamma \in \mathbb{R}^{(M-4(K_t-1)) \times C_o}$，它将输出映射到$Z \in \mathbb{R}^{n \times C_o}$上，而全连接层$\hat{v} = Zw + b$，其中$w \in \mathbb{R}^{C_o \times 1}$，$b \in \mathbb{R}^{C_o \times 1}$。因此，模型最后的输出结果$\hat{v} \in \mathbb{R}^n$。

而最后，模型的损失函数是预测值和真实值的距离度量：

$$L(\hat{v}; W_\theta) = \sum_t ||\hat{v}(v_{t-M+1}, \ldots, v_t, W_\theta) - v_{t+1}||^2,$$

其中，W_θ是所有可训练参数，$\hat{v}$是预测值，V_{t+1}是真实值。

残差连接：

作为一个小优化，$STGCN$模型在每个时域卷积块和空域卷积块中都使用了残差连接。

实验及结果：

1. 数据集

实验在北京市交通委员会和加州交通局收集的两个真实世界交通数据集 BJER4 和 PeMSD7 上验证了$STGCN$模型。每个数据集都包含交通观测值和地理信息的关键属性以及相应的时间戳。

2. 不同方法在各数据集上的性能比较

如下图所示，ARIMA（整合移动平均自回归模型）是典型的统计方法，由于不能处理复杂的时空图数据，表现最差；传统的机器学习方法由于误差累计、缺少空间信息效果也一般；$GCGRU$也是一种同时用到了时间和空间信息的方法，效

果与本文接近（在时域，本文使用CNN，$GCGRU$使用RNN）；而本文提出的$STGCN$效果最好；同时还试验了一阶切比雪夫GCN的效果，略差与$ChebGCN$。

Model	BJER4 (15/ 30/ 45 min)		
	MAE	MAPE (%)	RMSE
HA	5.21	14.64	7.56
LSVR	4.24/ 5.23/ 6.12	10.11/ 12.70/ 14.95	5.91/ 7.27/ 8.81
ARIMA	5.99/ 6.27/ 6.70	15.42/ 16.36/ 17.67	8.19/ 8.38/ 8.72
FNN	4.30/ 5.33/ 6.14	10.68/ 13.48/ 15.82	5.86/ 7.31/ 8.58
FC-LSTM	4.24/ 4.74/ 5.22	10.78/ 12.17/ 13.60	5.71/ 6.62/ 7.44
GCGRU	3.84/ 4.62/ 5.32	9.31/ 11.41/ 13.30	5.22/ 6.35/ 7.58
STGCN(Cheb)	**3.78/ 4.45/ 5.03**	**9.11/ 10.80/ 12.27**	**5.20/ 6.20/ 7.21**
STGCN(1^{st})	3.83/ 4.51/ 5.10	9.28/ 11.19/ 12.79	5.29/ 6.39/ 7.39

Table 1: Performance comparison of different approaches on the dataset BJER4.

Model	PeMSD7(M) (15/ 30/ 45 min)			PeMSD7(L) (15/ 30/ 45 min)		
	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE
HA	4.01	10.61	7.20	4.60	12.50	8.05
LSVR	2.50/ 3.63/ 4.54	5.81/ 8.88/ 11.50	4.55/ 6.67/ 8.28	2.69/ 3.85/ 4.79	6.27/ 9.48/ 12.42	4.88/ 7.10/ 8.72
ARIMA	5.55/ 5.86/ 6.27	12.92/ 13.94/ 15.20	9.00/ 9.13/ 9.38	5.50/ 5.87/ 6.30	12.30/ 13.54/ 14.85	8.63/ 8.96/ 9.39
FNN	2.74/ 4.02/ 5.04	6.38/ 9.72/ 12.38	4.75/ 6.98/ 8.58	2.74/ 3.92/ 4.78	7.11/ 10.89/ 13.56	4.87/ 7.02/ 8.46
FC-LSTM	3.57/ 3.94/ 4.16	8.60/ 9.55/ 10.10	6.20/ 7.03/ 7.51	4.38/ 4.51/ 4.66	11.10/ 11.41/ 11.69	7.68/ 7.94/ 8.20
GCGRU	2.37/ 3.31/ 4.01	5.54/ 8.06/ 9.99	4.21/ 5.96/ 7.13	2.48/ 3.43/ 4.12 *	5.76/ 8.45/ 10.51 *	4.40/ 6.25/ 7.49 *
STGCN(Cheb)	**2.25/ 3.03/ 3.57**	5.26/ **7.33/ 8.69**	**4.04/ 5.70/ 6.77**	**2.37/ 3.27/ 3.97**	**5.56/ 7.98/ 9.73**	**4.32/ 6.21/ 7.45**
STGCN(1^{st})	2.26/ 3.09/ 3.79	**5.24/** 7.39/ 9.12	4.07/ 5.77/ 7.03	2.40/ 3.31/ 4.01	5.63/ 8.21/ 10.12	4.38/ 6.43/ 7.81

Table 2: Performance comparison of different approaches on the dataset PeMSD7.

3. 早高峰和晚高峰时段的速度预测结果

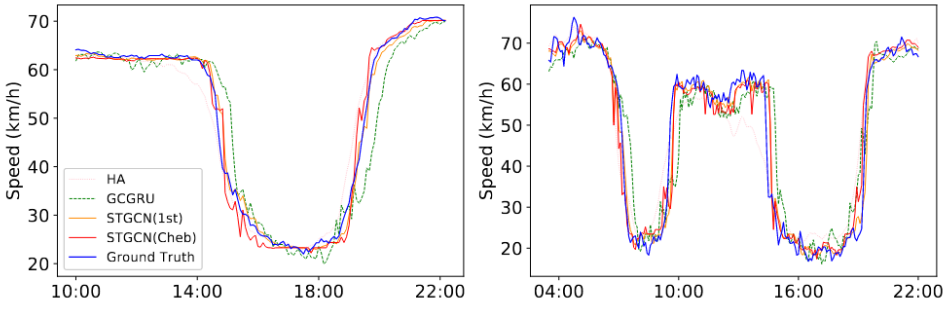


Figure 4: Speed prediction in the morning peak and evening rush hours of the dataset PeMSD7.

如上图所示，重点对比基于$GCGRU$、$STGCN(1st)$和$STGCN\ (Cheb)$可以看出在速度大幅变化时，$STGCN$比 GCGRU 要准确得多；在速度值最低点上，$STGCNA$的预测也更加准确。这是因为$STGCN$在时域上使用CNN，对历史数据的依赖比RNN更低。

4. 不同模型的训练时间比较

Dataset	Time Consumption (s)		
	STGCN(Cheb)	STGCN(1^{st})	GCGRU
PeMSD7(M)	**272.34**	271.18	3824.54
PeMSD7(L)	1926.81	**1554.37**	19511.92

Table 3: Time consumptions of training on the dataset PeMSD7.

如上表所示，在训练时间上$GCGRU$用时 3824s， $STGCN(Cheb)$ 用时 272s，而$STGCN(1st)$在$STGCN(Cheb)$上快 20%。

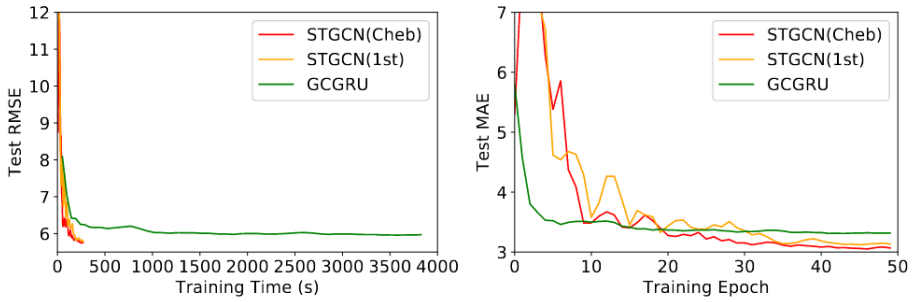


Figure 5: Test RMSE versus the training time (left); Test MAE versus the number of training epochs (right). (PeMSD7(M))

而上图还表明，$STGCN$模型不仅可以实现更快的训练过程，还更容易的收敛。由于$ST-Conv$模块的特殊设计，$STGCN$在平衡时间消耗和参数设置方面具有卓越的性能。（$STGCN$参数量是4.54×10^5，约为$GCGRU$的 2/3。）

总结：

$STGCN$是处理结构化时间序列的通用框架。它不仅能够解决交通网络建模和预测问题，还可以应用到更通用的时空序列学习任务之中。$STGCN$的时空卷积模块结合了图卷积和门控时序卷积，可以提取最有用的空间特征，并连贯地捕获最本质的时间特征。此外，最重要的是，由于该模型完全是由卷积结构组成的，因此，它能以更少的参数和更快的训练速度实现对输入的并行化处理。这种经济架构使该模型能够更高效地处理大规模网络。

对本文的感悟：

最为经典的一篇必读的时空图神经网络论文，具有非常的学习和借鉴意义。