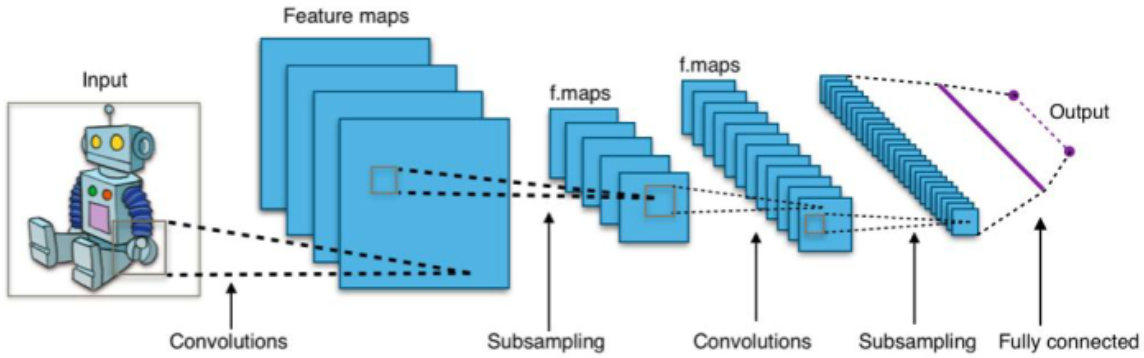
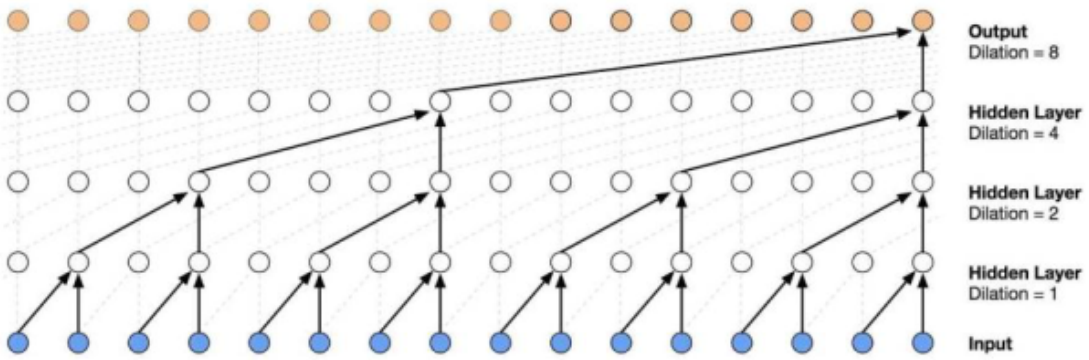
图神经网络

CNN工作

卷积神经网络(CNN)在欧式数据上，例如图片、文字、语音、视频取得了巨大的成功。比如在图像分类、对象检测、机器翻译等任务取得了显著的进展。





CNN的成功主要来自于：

通过局部化的卷积核学到局部稳定的结构，通过层级堆叠，将学到的局部的结构变成层次化的、多尺度的模式。

卷积核具有平移不变性:

用卷积核学到的特征与位置无关，在空间上有一个平移不变性。

如何把定义在欧式数据上的卷积神经网络迁移到非欧式数据上来。非欧式数据，非规则类数据，比如图数据。

在图上定义卷积最大的困难是如何定义卷积概念?

早期的图卷积工作主要有两方面：

- 如何定义图上的卷积
- 如何定义图上的池化

连续型卷积

h(t)信号比原来的f(t)信号更加平滑

$$h(t) = (f * g)(t) \overset{\text{def}}{=} \int f(t)g(t-\tau)\,d\tau$$

离散性卷积

更多是在离散情况下的卷积操作

$$\begin{aligned} & h(x,y) \\ & = (f * g)(x,y) \\ & \overset{\text{def}}{=} \sum_{m,n} f(x-m, y-n)g(m,n) \end{aligned}$$

卷积更像是一个图上信号处理的一个过程

对于卷积来说，一开始就有两种定义方法，一种是谱方法，一种是空间方法。

谱方法是空间方法的一个子集。

现有的定义卷积的方法

谱方法

谱方法是空间方法的一个子集，谱方法是试图在谱域范围内定义卷积。

在图上不满足平移不变性，所以无法在图上直接定义节点的卷积。为此，把图上的信号变到谱域，在谱域上定义卷积，然后再变换到图上。

空间方法

卷积被定义为位于目标顶点邻域的加权平均后的顶点。面临的主要困难是，每个节点的领域大小不一样，参数共享实现比较困难。

谱方法

给定一个图$G = (V, E, W)$, V 是节点集，$n = |V|$,E是边集合，$W \in R^{n \times n}$是边的权重集合

每一个节点有d个特征，于是就有特征矩阵$X \in R^{n \times d}$，其中X中的每一列可以看作是图上节点的信号。

图上的拉普拉斯

图上的拉普拉斯是对图进行谱分析的一个基本工具，刻画了信号在图上的平滑程度。

$L = D - W$ 这里 $D_{ii} = \sum_j W_{ij}$

拉普拉斯矩阵一个标准化的版本;

$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ 这里的I是单位矩阵。

图上的傅里叶变换

图上的信号转换到谱域，在谱域处理完以后，在变换到节点域。

如果一个图有n个节点，图上的信号就是一个n维向量。我们需要把这个n维向量变换到一个新的域，需要一组基底。这个基底就是拉普拉斯矩阵n个特征向量。这些特征向量是正交的。

拉普拉斯矩阵的特征分解：

$L = U \Lambda U^T$

这里$U = [u_1, \dots u_n]$ ， $\Lambda = [\lambda_1, \dots \lambda_n]$

图上的傅里叶变换:

图上的信号$x \in R^n$

$\hat{x} = U^T x$

图傅里叶反变换

$x = U \hat{x}$

在谱域上的卷积定义

两个信号的卷积可以看作：是它们傅里叶变换后的点积。

输入图信号x，卷积核y，先将他们变换到谱域，在谱域中进行点积，最后将点积后的结果反变换到空域，过程如下公式所示:

$x *_G y = U((U^T x) \bigodot (U^T y))$

以上公式揭示了在谱域上的卷积定义。

进一步对公式讨论：

$U^T y = [\theta_0, \dots, \theta n - 1]^T$

$g_\theta = diag([\theta_0, \dots, \theta n - 1])$

那可以将公式

$x *_G y = U((U^T x) \bigodot (U^T y))$

变成

$x *_G y = U g_\theta U^T x$

步骤可以分为:

1. 图上的傅里叶变换$U^T x$
2. 对变换后的信号做卷积$g_\theta U^T x$
3. 傅里叶逆变换$U g_\theta U^T x$，转换到节点域

谱域上的CNN：

$j = 1, \ldots, f_{k+1}$

$x_{k+1,j} = h(\sum_{i=1}^{f_k} UF_{k,i,j}U^T x_{k,i})$

其中$x_{k,i}$ 表示了第k层的信号

$F_{k,i,j}$表示第k层的卷积核

特点：

- 依赖于拉普拉斯矩阵的特征分解，复杂度很高
- U是一个稠密矩阵，乘以x的时候时间复杂度很高
- 不是局部的，在节点域中进行处理的时候，对节点的影响不是来自于节点的领域，而是来自于所有的节点（not localized）， 这不是我们想看到的卷积的操作

ChebyNet 卷积核参数化

卷积核参数多项式近似

$g_\theta = diag([\theta_0, \ldots, \theta n - 1])$ 是一个自由的卷积核，有n个参数的矩阵。

用$g_\beta(\Lambda) = \sum_{k=0}^{k-1} \beta_k \Lambda^k$ 其中$\Lambda = diag(\lambda_1, \ldots \lambda n)$

ChebyNet:

$x *_G y = U g_\theta U^T x = \sum_{k=0}^{k-1} \beta_k L^k x$

- 自由的参数个数从n个变成了k个
- 不再显示地依赖于U
- L是一个稀疏矩阵并且是局部化的矩阵

Graph Wavelet Neural Network

用小波变换基底代替傅里叶变化基底

傅里叶变换	**小波变换**
正弦波	局部有信号，其他地方都是0
稠密	稀疏
Not localized	Localized
High Computational cost	Low Computational cost
基底表示:U	基底表示:$\Psi_s = Ue^{\lambda s}U^T$

小波变换基底可以基底的角度解决傅里叶变换很多没解决的问题

把基底换成小波变换基底

图小波正变换

$x^* = \Psi_s^{-1} x$

图小波逆变换

$x = \Psi_s x^*$

利用小波变换实现图卷积的公式表达:

$x *_G y = \Psi_s((\Psi_s^{-1}x) \bigodot (\Psi_s^{-1}y))$

图小波神经网络

$x_{k+1,j} = h(\sum_{i=1}^{p} \Psi_s F_{k,i,j} \Psi_s^{-1} x_{k,i})$

$j = 1, \dots q$

参数个数:$O(n * p * q)$

n:节点个数

p:输入信号

q:输出信号维度

这样的复杂度还是不能接受的，在复杂的社交网络中，复杂度还是过高了。

有新的观点认为，图上的卷积和图上的特征变换不应该放在一起进行，不然参数量会非常大

$x_{k+1,j} = h(\sum_{i=1}^{p} \Psi_s F_{k,i,j} \Psi_s^{-1} x_{k,i})\ j = 1, \dots q$

特征变换:

$y_{k,j} = \sum_{i=1}^{p} T_{j,i} x_{k,i}\ \ T \in^{q \times p}$ 有 $p * q$个参数

图卷积过程:

$x_{k+1,j} = h(\Psi_s F_k \Psi_s^{-1} y_{k,j})$ F_k是一个由n个 参数的对角矩阵

参数个数变为$O(n + p * q)$

空间方法

Learning Convolutional Neural Networks for Graphs.

2016工作

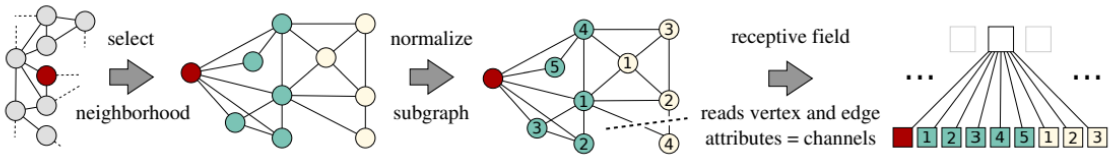
这个工作试图用类比的方法，把CNN迁移到graph上。

卷积有三步骤：

1. 选定领域节点
2. 给领域节点编号
3. 参数共享

这三步里只有第一步需要做平移不变性。

- 对每一个节点，选定固定个数的节点作为它的邻居,根据邻近度的度量
- 选择固定窗口大小的节点，进行参数共享



GraphSAGE

2017年

要选一个固定个数的领域，不应该用邻近度矩阵

以自己为中心，随机游走，选择固定个数的节点。离自己越近的节点被选择的概率越大。选完之后聚合。

这里的卷积主要体现在aggregate和 combine 上。

GCN

2017年

论文里说是GCN是ChebyNet一个简化的版本。但是他应该是谱方法的一个特例，又是空间方法的一个起点。这个方法也是很多人熟悉的图卷积算法。

$Z = f(X, A) = softmax(\hat{A}ReLU(\hat{A}XW^{(0)})W^{(1)})$ 其中的$W^{(0)}$和$W^{(1)}$用于特征变换

所谓的卷积操作就是，邻居节点特征变换之后，做一个加权平均。权重直接用拉普拉斯矩阵直接定义。

GAT

Graph Attention Network 2018年

认为：

GCN没有卷积，参数共享也没有，参数传递靠的是拉普拉斯矩阵，参数共享只是在特征变换之后做了一个参数共享。

- 学习聚合矩阵，即GCN中的拉普拉斯矩阵，通过注意力机制。
- 参数共享分为两部分。一部分特征变换，另一部分，注意力机制。

卷积核就是attention 参数 a

$$\alpha_{ij} = \frac{exp(LeakyReLU(\vec{a}^T[W\vec{h}_i||W\vec{h}_j]))}{\sum_{k\in N_i} exp(LeakyReLU(\vec{a}^T[W\vec{h}_i||W\vec{h}_k]))}$$

GAT不是一个简单的均值聚合，是一个关注邻居特征的不同的聚合函数

MoNet

更通用、更一般的空间方法

图上的核函数，定义图上两个节点之间相似度的方式。所谓的卷积就是对这些不同相似度的定义方式的加权平均。卷积核的参数就是核函数的权重。

$(f * g)(x) = \sum_{j=1}^{J} g_j D_j f$ 其中的g_j就是卷积核

核函数在谱方法里面就是变换的基底。

在空间方法里面就是选择哪些邻居和我的相似度是怎么样的为基础。

谱方法和空间方法具体有什么联系？

Heat Kernel

2019年

谱方法是空间方法的特例

$(f*g)(x)=\sum_{j=1}^{J} g_j D_j(x) f$

卷积核函数：节点间相似性、距离

谱方法需要显示地定义卷积核，空间方法不需要显示定义。谱方法需要知道到底把节点投影到哪个空间里去

Spectral CNN

$y=Ug_\theta U^T x=(\theta_1 u_1 u_1^T+\theta_2 u_2 u_2^T+\theta_3 u_3 u_3^T)x$

ChebyNet

$y=(\theta_0 I+\theta_1 L+\theta_2 L^2+\ldots+\theta_{K-1}L^{K-1})x$

GCN

$y=\theta(I-L)x$

节点分类任务对图上信号平滑性有一定的要求。

$x^T Lx$刻画了信号的平滑程度

$x^T Lx=\sum_{(u,v\in E)}A_{uv}(\frac{x_u}{\sqrt{d_u}}-\frac{x_v}{\sqrt{d_v}})^2$

如果信号正好是一个特征向量。

$\lambda_i=u_i^T Lu_i$ 信号的平滑程度就是这个特征向量关于图的特征值

λ是特征值 u是特征向量

L有一个退化的特征向量，乘以L 结果是0 这是平滑性最好的信号

$u_i u_i^T(1\le i\le n)$是一个基本的滤波器，只能让频率是λ_i的信号通过滤波器

每一个卷积操作，都是基础滤波器的组合

$x=\alpha_1 u_1+\alpha_2 u_2 \alpha_3 u_3+\ldots \alpha_n u_n$

$u_i u_i^T x=\alpha_i u_i$

标准谱方法就是基础滤波器的线性组合

$\theta_1 u_1 u_1^T+\theta_2 u_2 u_2^T+\ldots \theta_n u_n u_n^T$

L^k可以看作是这组滤波器带上系数$\{\lambda_i^k\}_{i=1}^n$

频率越高，系数越高，所以L^k是一个高通滤波器 但是高频信号，在节点分类里不体现平滑性

为什么GCN反而比ChebyNet好

GCN只考虑了k=0,k=1,避免高频信号带来的影响

直接设计一个低通滤波器

$\{e^{-skl}\}$ s是缩放参数，k是序号

$e^{-sL} = Ue^{-s\Lambda}u^T\Lambda = diag(\lambda_1, \lambda_2, \dots \lambda_n)$

可以压制高频信号

这也是空间领域的选择问题，这也是空间方法的根本问题。

图上的池化

图上的节点级别对应了图像的像素级别，是不需要池化的

早期的任务是不需要池化，只需要卷积就够了。

Graph coarsening

图粗化 -> 也可以理解为下采样

- 图上的节点先聚类，然后每一个类别在形成一个超级节点 这样就可以让图的大小越来越小 最后变成一个节点
- 图上 节点的合并可以实现合并(先验性) 也可以在训练过程中进行合并（DiffPooling)

  每个节点属于一个cluster的概率

Node Selection

学习一个矩阵，这个矩阵表征了节点的重要程度。根据这个节点重要性矩阵来选择一部分节点。