Graph Attention Networks 学习笔记

大纲

背景

目的与动机

本篇文章的目的是，应用注意力机制到图相关的机器学习任务中，以构建一个可以应用于图上的注意力模型。

现实中有很多数据是以图的结构存在的。比如三维网孔，社交网络，通讯网络，生物网络或者是神经元的连接。用传统的机器学习方式在这些图结构的数据上取得的效果并不理想。可是图结构的信息又具有很高的价值，有分析研究的意义。所以，包括本篇论文在内的很多研究都致力于找到一个合适强大的模型，可以提取表达图结构中的数据。

本文中提出的图注意力网络 *(Graph Attention Network)* ，改善了在这一领域许多已有模型的不足之处，并能够很好的应用到直推式学习 *(transductive)* 与归纳式学习 *(inductive learning)* 中，都取得了很好的效果。

已有的有关工作

在图学习领域已有了很多的尝试与研究工作。早期的工作有递归神经网络，将传统的神经网络拓展到了图上。以及随后引入的图神经网络 *(GNN)* 。

人们一直在尝试把"卷积"也推广到图上，推广"卷积"的研究工作大体可以分为两个方向，一个是谱图的方法，一个是非谱图的方法。

本篇论文中提出的图注意力网络改善了已有工作的许多不足之处，解决了许多之前的工作难以解决的问题，比如：

1. 利用图谱的模型中很多都需要耗费大量计算资源的计算密集型操作，比如矩阵求逆，但是图注意力网络避免了这一点，更加高效，也可以进行并行计算
2. 已有的模型很多依赖于提前知道图的完整结构，但图注意力网络不要求这一点
3. 图注意力网络可以给节点的邻居分配不同的权重，但是先前的很多模型，比如Kipf提出的GCN做不到这一点
4. 非谱图方向上的很多已有的模型，都不能很好地处理节点度数不同的问题。而借助于注意力机制，图注意力网络可以很好的处理任意度数的节点

模型引入

根据论文，设计这个模型的初衷是受了序列处理任务中注意力机制巨大作用的启发。

注意力机制最初被引入的目的是解决Seq2Seq模型的一些问题，主要为当时已有的Seq2Seq模型试图将输入语句的所有信息全部压缩到一个向量中，很大程度上限制了模型的表达能力；而且将语句中所有的词语平等对待，没有利用不同词语在句中的重要性不同这一特点，并不合理。注意力机制的引入使得Seq2Seq模型可以学习到语句中的更多信息，并且可以分析语句中每个单词的重要程度，大大提升了Seq2Seq模型的效果。

随后，注意力机制被广泛应用到了序列模型或是图像识别领域，在各方面都取得了重要的成果。

注意力机制脱胎于对人"注意力"的理解。人类在生活中，周围环境变化多端，信息纷乱庞杂，但无论在什么任务中，人类都只会关注相对一小部分的信息，也就是所谓的"注意"，专注于最重要的地方，提取出更多有价值的信息。

本文提出的图注意力模型将注意力机制引入到了图模型中。图注意力模型利用自注意力 *(self-attention)* 策略，根据节点的邻居计算出图中节点的隐藏表示。论文中提到，注意力结构具有如下优点：

1. 注意力机制的计算是高效的，因为这些操作可以在不同相邻点对上并行。
2. 可以通过给邻居节点分配不同权重的方式应用到任意度数的节点上。
3. 模型可以直接被应用到归纳式学习任务中，并可以泛化到完全不可见的图上。

模型介绍

Graph Attentional Layer

图注意力层 *(Graph Attentional Layer)* ，是构建图注意力网络的基本组成部分。事实上，图注意力网络就是由图注意力层堆叠而成的。

图注意力层的输入为节点的特征，

$$\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \cdots, \vec{h}_N\}, \mathbf{h} \in \mathbb{R}^{F \times N}$$

其中，F为特征向量的维数，N为图中节点数量。

图注意力层的输出为利用注意力机制计算出的节点的新的特征表达，

$$\mathbf{h}^{'} = \{\vec{h}_1, \vec{h}_2, \cdots, \vec{h}_N\}, \mathbf{h}^{'} \in \mathbb{R}^{F^{'} \times N}$$

其中$F^{'}$为输出特征向量的维度。

图注意力层的输出以这种方式计算：

$$\vec{h}_i^{'} = \sigma(\sum_{j \in N_i} \alpha_{ij} W \vec{h}_j)$$

其中α即为注意力系数，$W \in \mathbb{R}^{F^{'} \times F}$为系数矩阵，$\sigma$为激活函数，$N_i$为节点$i$所有邻居节点的集合。

注意力系数的计算

所以，模型的重点就在于注意力系数的计算，要计算注意力系数，我们先需要计算自注意力系数，事实上：

$$\alpha_{ij} = softmax_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})}$$

而我们有这样的注意力机制来计算注意力系数：

$$e_{ij} = a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$$

其中a即为注意力机制。

> 在注意力机制的定义里，并没有对i与j做任何的限定，这也意味着图中任意一个j都能对i产生影响，这意味着我们丢弃了所有的图结构信息，显然是不合理的。
>
> 为了在注意力机制中利用图结构的信息，图注意力模型利用了掩盖的注意力 *(masked attention)* 计算，也就是说，只计算节点i一阶邻居的注意力系数。

所以，现在我们只要知道a就能拼上图注意力层的最后一块拼图了，a在图注意力模型中是一个单层前馈网络：

$$a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j) = LeakyReLU(\vec{a}^T[\mathbf{W}\vec{h}_i || \mathbf{W}\vec{h}_j])$$

其中，$\vec{a}^T \in \mathbb{R}^{2F^{'}}$为系数向量。

由此，我们可以得到注意力系数的最终表达式：

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

Exponential Linear Unit (ELU)的使用

在论文对模型的介绍中，没有明确的指出图注意力层中激活函数σ的选取。但在后面的实验评估中提到他们在实验中选择了ELU，Exponential Linear Unit。

ELU的表达式：

$$\sigma(x) = \begin{cases} x & x > 0 \\ \alpha(e^x - 1) & x \leq 0 \end{cases}$$

ELU在很多模型中应用都带来了更好的效果与更快的训练。

Multi-head Attention

论文中提到，为了稳定模型自注意力的学习过程，图注意力网络采用了multi-head attention。

顾名思义，multi-head attention的含义为采用与训练多个注意力机制a。一般来说，multi-head attention的作用是，让模型可以从更多种角度学习注意力分配的模式，从而获得更好、更稳定的效果。

引入multi-head attention后，模型的输出表达为

$$\vec{h}_j^{'} = ||_{k=1}^{K} \sigma(\sum_{j \in N_i} \alpha_{ij}^k W^k \vec{h}_j)$$

其中K为设置的不同注意力机制的数量，$||$代表向量的拼接。

因此，输出特征向量的维度就由F'变成了KF'。

特别地，对于最后一层图注意力网络，我们这样计算输出：

$$\vec{h}_j^{'} = \sigma(\frac{1}{K} \sum_{k=1}^{K} \sum_{j \in N_i} \alpha_{ij}^k W^k \vec{h}_j)$$

由于最后一层的输出即为网络的输出，因而对于输出维度有要求(分类任务)，故向量拼接将不再具有意义，所以这里将拼接改为了平均，并将激活函数移到了外面。

模型评价

优势

在第一部分中，我提及了图注意力网络试图解决的问题，以及相较于已有模型的改善，其中已经包含了有关图注意力网络优势的内容，可以分为如下几个方面：

1. 更高的计算效率。

   注意力机制的计算是可以在所有边之间并行的，并且最终向量表达的计算也可以在所有节点间并行。不仅如此，注意力机制的计算并不涉及复杂耗时的矩阵运算，比如特征值分解。而且，multi-head attention中的不同机制也可以并行计算。

2. 更强的表达能力。

   相较于GCN，图注意力网络能够通过注意力机制，给邻居节点分配不同的权重，这会大大增强模型的能力。不仅如此，借助注意力机制学习出来的权重分配策略，图注意力网络将拥有更好的解释性。

3. 注意力机制使模型能够轻松处理不同度数的节点。

GraphSAGE依赖于在节点的邻居中取样固定个数的节点来保证一致性，并且在使用基于LSTM的neighbourhood aggregator时能够取得比较好的效果——这需要一个一致的相邻节点的先后顺序才行得通。为了解决这个问题，GraphSAGE向LSTM中输入随机顺序的节点序列。利用注意力机制，图注意力模型完全没有这些问题，可以在所有相邻节点中提取信息。

4. 应用更为广泛。

   还是因为注意力机制带来的好处，图注意力模型可以应用到：

   - 有向图中，此时，只需要忽略一条边上某个节点对另一个节点的影响即可
   - 归纳式学习任务，甚至是在训练中完全不可见的图

表现

一如论文中所呈现的那样，图注意力网络在直推式学习与归纳式学习任务中都取得了媲美甚至超过state-of-the-art水准的表现。

Transductive

Method	Cora	Citeseer	Pubmed
MLP	55.1%	46.5%	71.4%
ManiReg (Belkin et al., 2006)	59.5%	60.1%	70.7%
SemiEmb (Weston et al., 2012)	59.0%	59.6%	71.7%
LP (Zhu et al., 2003)	68.0%	45.3%	63.0%
DeepWalk (Perozzi et al., 2014)	67.2%	43.2%	65.3%
ICA (Lu & Getoor, 2003)	75.1%	69.1%	73.9%
Planetoid (Yang et al., 2016)	75.7%	64.7%	77.2%
Chebyshev (Defferrard et al., 2016)	81.2%	69.8%	74.4%
GCN (Kipf & Welling, 2017)	81.5%	70.3%	**79.0%**
MoNet (Monti et al., 2016)	81.7 ± 0.5%	—	78.8 ± 0.3%
GCN-64*	81.4 ± 0.5%	70.9 ± 0.5%	**79.0** ± 0.3%
GAT (ours)	**83.0** ± 0.7%	**72.5** ± 0.7%	**79.0** ± 0.3%

Inductive

Method	PPI
Random	0.396
MLP	0.422
GraphSAGE-GCN (Hamilton et al., 2017)	0.500
GraphSAGE-mean (Hamilton et al., 2017)	0.598
GraphSAGE-LSTM (Hamilton et al., 2017)	0.612
GraphSAGE-pool (Hamilton et al., 2017)	0.600
GraphSAGE*	0.768
Const-GAT (ours)	0.934 ± 0.006
GAT (ours)	**0.973** ± 0.002

不足与改进

在论文中，作者也提及了图注意力网络的不足与可以改进之处：

1. 由于实现所用框架的限制(TensorFlow)，张量的稀疏运算只能对二阶张量进行，限制了模型的Batch性能。
2. 模型的感知野 *(receptive field)* 受限于网络的深度。可以利用残差连接来加深模型的深度，来获取更好的性能。
3. 在边上分布式的并行计算可能会导致大量的重复计算。
4. 可以利用注意力机制，对模型的解释性进行深入研究。
5. 可以拓展该方法，使模型能够对图而不是节点进行分类任务。
6. 可以通过引入边的特征，来进一步拓展模型。

感想

总的来说，图注意力模型是一个很棒的模型。

注意力机制在图上的应用是合适且符合直觉的。一方面，图结构的多变性，特别是节点度数的任意性，是应用于图的模型必须要面对的一个问题。GNN中的aggregation function是一个较为直观的解决办法：将相邻节点的特征向量求和。可以说，注意力机制其实也是某种意义上的aggregation function，只是注意力机制更为强大，他并不等同地处理每一个节点，或者以事先规定的某种机制分配权重，而是让模型在训练过程中自己学习权重的分配策略。另一方面，图中节点受相邻节点影响的程度，就如同句子中某个单词与其他单词的关联一样，必然存在一些有迹可循的特征。比如，在如Cora这样的论文引用关系数据集中，字典中有更多相似单词的论文之间也许会具有更强的关联性。利用注意力机制发掘出这些隐藏的关联性，直觉上是有益于模型的表现的。

注意力机制在图注意力网络中的成功应用，也预示着图模型处理与自然语言处理之间的紧密联系。从Deepwalk到GAT，应用在NLP中的技术很多时候能类似地应用到图模型中，取得很好的效果，或至少给出有价值的参考与启发。在Deepwalk的论文中提及，代表社交关系的图数据很大程度上与自然语言数据类似，他们都符合幂律，满足幂律分布。注意力机制在NLP领域发挥了重要作用，图注意力网络也成功地将注意力机制应用到了图中。

图模型领域的研究成果目前似乎还没有如图像识别或自然语言处理领域的那么丰富深入，可能也是因为图是比图像与语言更为抽象的一种模型。图像识别模型可以模仿或参考人类识别图形信息的方法，自然语言处理模型也难参考人类对语句的理解方法，但是基本没有人对图的理解方式可供参考。如果图处理领域能够从自然语言处理领域得到很好的参考与指引，那的确是一件大有裨益的事情。

在以后的研究中，可以尝试优化注意力机制的计算方式，尝试采用其他的计算方式。并且拓展注意力机制，将边的特征也考虑在内。