GraphSAGE理论与实践

# 1 前言

GraphSAGE是2017年提出的一种图神经网络算法，论文题目是《Inductive Representation Learning on Large Graphs》。GraphSAGE取自Graph SAmple and aggreGatE，SAmple指如何对邻居个数进行采样；aggreGatE指拿到邻居的embedding之后如何汇聚这些embedding以更新自己的embedding信息。GraphSAGE解决了GCN网络的局限性: GCN训练时需要用到整个图的邻接矩阵，依赖于具体的图结构，一般只能用在直推式学习（Transductive Learning）。GraphSAGE使用多层聚合函数，每一层聚合函数会将节点及其邻居的信息聚合在一起得到下一层的特征向量，GraphSAGE采用了节点的邻域信息，不依赖于全局的图结构。在工业界一直受到重视。

图神经网络的任务一般有Transductive (直推式)和Inductive(归纳式)。Transductive通常指要预测的节点在训练时已经出现过，例如有一个作者关系网络，知道部分作者的类别，用整个网络训练GCN，最后预测未知类别的作者。Inductive指要预测的节点在训练时没有出现，例如用今天的图结构训练，预测明天的图。

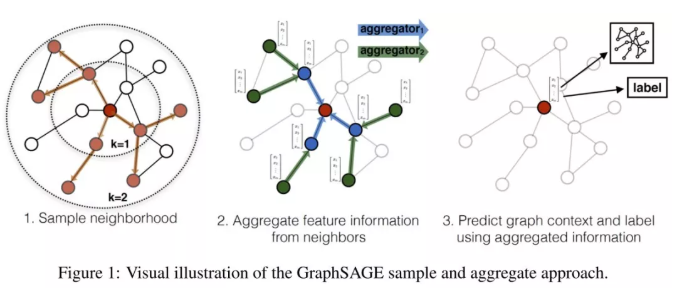
在GraphSAGE之前，图神经网络主要的方法有DeepWalk，GCN等，这些GNN方法的不足在于需要对全图进行学习，而且是以Transductive learning为主，也就是说需要在训练的时候，图就已经包含了要预测的节点。训练时不包含的节点，预测时也无法给出预测结果。

在实际应用中，图的结构会频繁变化，在预测阶段，可能会往图中新添加新节点。针对这种情形，采用直推式的模型（比如DeepWalk，GCN等就无法满足了）。GraphSAGE就是针对这种场景提出的，GraphSAGE = Graph Sample Aggregate，即对图进行采样（sample）和聚合（aggregate）。

# 2 GraphSAGE

## 2.1 GraphSAGE基本思路

GraphSAGE 包含采样和聚合 (Sample and aggregate)，首先使用节点之间连接信息，对邻居进行采样，然后通过多层聚合函数不断地将相邻节点的信息融合在一起。用融合后的信息预测节点标签。下图展示了 GraphSAGE 的聚合过程，采用了两层聚合层。



以上图为例进行说明GraphSAGE采样和聚合的过程：

在第一幅图是采样（sample）的过程。在这个图中，需要对最中心的节点进行embedding更新，先从与它相邻的节点中选择S1个（图中选择了3个）节点。这里假设k=2，那么我们还需要对第二层再进行采样，也就是对刚才选择的S1个邻居节点，再选择他们的邻居节点。注意：要使每个节点采样的邻居数保持一致，这样才可以把多个节点及他们的邻居拼成Tensor进行训练。（实验表明最佳参数为迭代次数K=2，两次迭代所选邻居数S1\*S2<=5。作者实验中S1 = 25，S2 = 10）

在第二幅图中进行aggregate的过程，也就是先拿邻居节点的邻居更新邻居的信息，再用更新后的邻居的信息来更新目标节点（也就是最中心的节点）的信息。

在第三幅图中进行节点预测，假如要预测一个未知节点的信息，只需要用它的邻居节点的信息来预测即可。

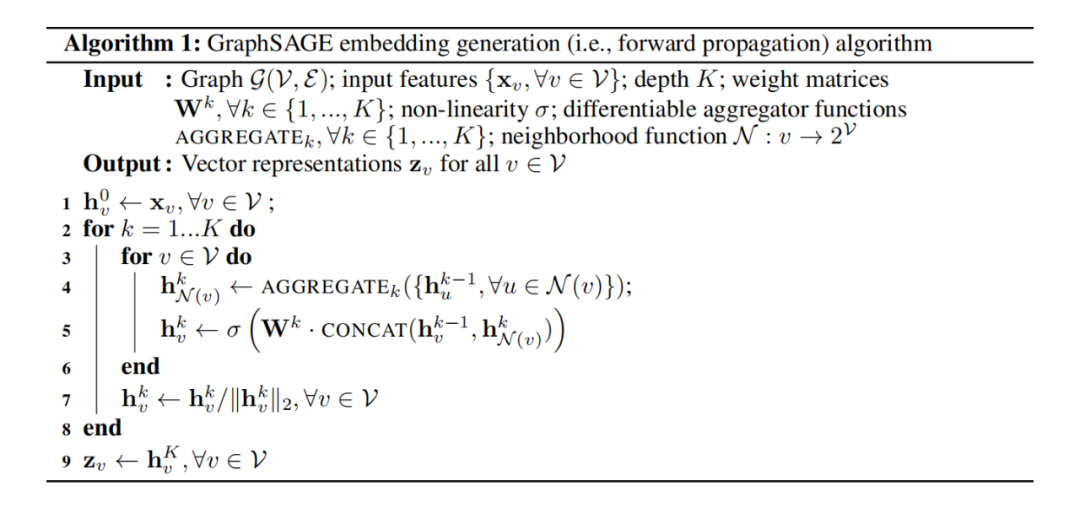
举例：如果我想知道小明是一个什么性格的人，可以找几个他关系好的小伙伴观察一下，为了进一步确认，再选择他的小伙伴们的其他小伙伴，再观察一下。也就是通过小明的小伙伴们的小伙伴，来判断小明的小伙伴们是哪一类人，然后再根据他的小伙伴们，就可以粗略的得知，小明是哪一类性格的人了。

## 2.2 GraphSAGE算法流程

总体来说，GraphSAGE生成（更新）节点embedding的过程其实是假设已经完成了GraphSAGE的训练，因此模型的所有参数都已知了。

具体来说，这些参数包含K个聚合器AGGREGATEk（见下图第4行）中的参数，这些聚合器被用于将邻居节点的embedding信息聚合到节点上，以及一系列的权重矩阵W（见下图第5行），这些权值矩阵被用作在模型层与层之间传播embedding的时候做非线性变换。

下面是GraphSAGE产生embedding的过程，也被称为前向传播过程：

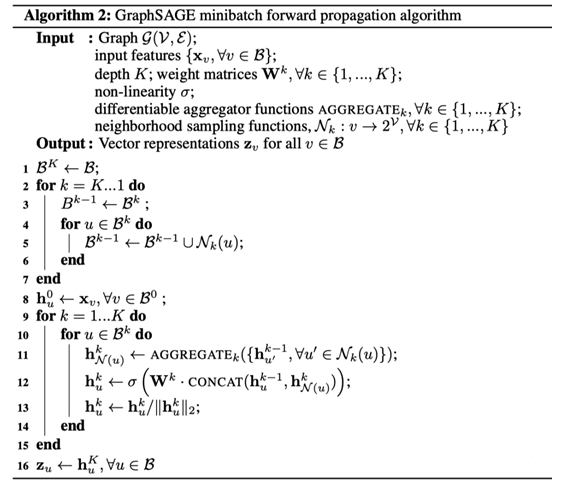


伪代码中的h0表示节点v的初始特征向量，包含K层聚合操作。在第k次聚合生成v节点特征向量时，会采用聚合函数把v节点的邻居信息融合在一起。

* 第一行是要计算的节点的初始特征输入；
* 第二行是第一个 for 循环遍历深度，可以理解为神经网络的层数；
* 第三行是第二个 for 循环是遍历图中所有节点；
* 第四行是从上一层神经网络中利用聚合函数聚合当前节点邻居的特征；
* 第五行是将当前节点的特征和邻居特征拼接并经过一个全连接网络得到当前节点的新特征；
* 第七行是归一化；
* 第八行是通过K层GCN后进行输出。

简单来说就是用k-1层的节点的邻居信息和自身信息来更新k层的节点信息。这里的K是聚合器的数量，也是权重矩阵的数量，还是网络的层数。网络的层数可以理解为需要最大访问到的邻居的跳数。比如在Figure1中，红色节点拿到了距离它一、二跳的邻居信息，那么网络的层数就是2。

但这样会出现一个问题：如果只是想计算某个新的节点的Embedding，其实没有必要把整张图的节点的Embedding都更新一遍。针对这个问题，作者给出了算法二：



上面的伪代码中，B = BK为要生成向量的节点集合，Bk-1是深度为1的邻域，B0为深度为K的邻域，B0包含的节点最多。Nk(u)表示u节点在第k层进行聚合时的邻域，节点在每一层的邻域数量都不同，通过采样得到。这里邻居采样的大小是固定的，以保证每个批处理单元大小都是固定的。

算法的第1到7行，其实就是一个sample的过程，并且将sample的结果保存到B中。接下来的9-15行，就是一个aggregate的过程，按照前面sample的结果，将对应的邻居信息 aggregate 到目标节点上来。

仔细观察，可以发现sample的过程是从K到1的（看第2行），而aggregate的过程是从1到K的（第9行）。这是因为采样的时候，先从整张图选择目标节点，然后对这些节点的邻居进行采样，并且逐渐采样到远一点的邻居上。而在聚合时，是先从最远处的邻居上开始聚合，最后第K层的时候，才能聚合到目标节点上来。

在图神经网络中，数据集都通常非常大，首先mini batch的思路非常重要，然后在GraphSAG中，只需要对自己采样的数据进行聚合，无需考虑其它节点。每个batch可以是一批sample结果的组合。这是GraphSAGE的精妙之处之一。

## 2.3 GraphSAGE聚合函数

在GraphSAGE中，聚合函数非常重要。关于聚合函数的选择有两个条件：

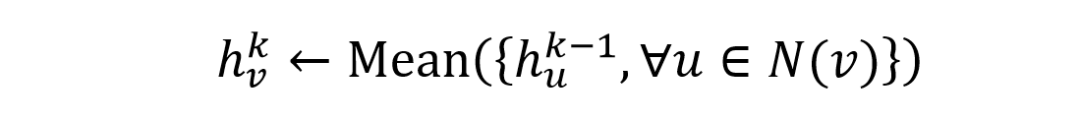
* 可导，因为要反向传递来训练目标的聚合函数参数；
* 对称，这里的对称指的是对输入不敏感，因为在聚合的时候，图中的节点关系并没有顺序上的特征。

所以在作者原文中选择的都是诸如Mean，max pooling之类的聚合器，虽然作者也使用了LSTM，但是在输入前会将节点进行shuffle操作，也就是说LSTM从序列顺序中并不能学到什么知识。

GraphSAGE提供了四种聚合节点的函数，其中Mean Aggregator是实验中效果最好的聚合器。

### 2.3.1 Mean aggregator

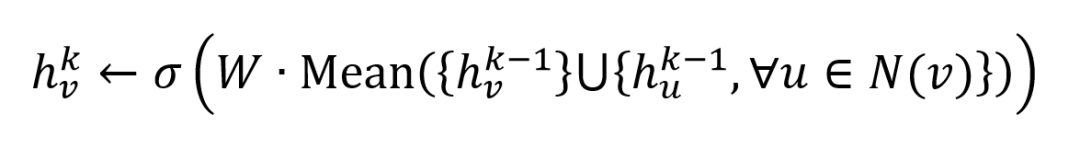
Mean aggregator，均值聚合器，对邻居节点和当前节点embedding向量取均值。



举个简单例子：比如一个节点的3个邻居的embedding分别是[1,2,3,4], [2,3,4,5], [3,4,5,6]，按照每一维分别求均值就得到了聚合后的邻居embedding为[2,3,4,5]。

### 2.3.2 GCN aggregator

GCN aggregator: 采用了类似GCN卷积的方式进行聚合，公式和Mean aggregator类似：



### 2.3.3 LSTM aggregator

LSTM aggregator，作者认为LSTM有比较好的抽取特征能力，因此也使用了LSTM进行聚合，但是因为LSTM是对称的，为了解决这个问题需要将LSTM应用于节点的邻居随机排序，即可使LSTM适应无序集合。

### 2.3.4 Pooling aggregator

Pooling aggregator，池化聚合器。这种聚合方式可以使得节点的每个邻居的Embedding 向量都可以独立的通过全连接的神经网络，通过这样的转换后最大池化操作可以聚合整个邻居集合。

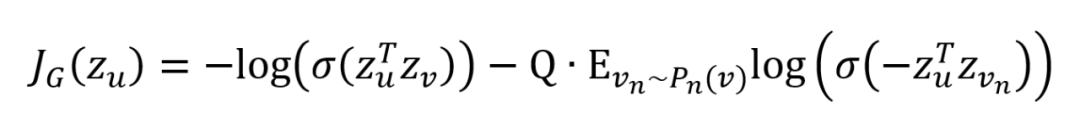


## 2.4 GraphSAGE训练

GraphSAGE训练的过程就是学习聚合器参数和权重变量W的过程。

GraphSAGE可以采用无监督训练或者有监督训练。其实不论选择哪种训练方式，我们的目的还是使用它来完成节点embedding的过程。

无监督训练采用负采样算法。也就是利用图中的邻居关系，当两个节点距离相近时默认这两个节点的embedding也相似；相反，如果两个节点的相距较远，那么它们的embedding应该差异很大。基于此，采用无监督训练时损失函数公式如下：



公式中的zu是经过GraphSAGE聚合之后的特征向量（也就是预测的embedding），节点v是节点u的邻居（这里邻居是广义的，比如说如果节点v和u在一个订场的随机游走中可达，那么也可以认为他们相邻），Pn是负采样分布，v\_n表示不是节点u的邻居，Q表示负采样的样本数量。负采样是指长全部节点中选择一批不是节点v邻居的节点作为负样本。因此，上面的损失函数的含义是：相邻节点的embedding的相似度尽可能大的情况下，保证不相邻的节点的embedding的期望相似度尽可能小。

在论文中，作者对邻居的定义是：直接使用DeepWalk进行随机游走，步长为 5，测试50次，走得到的都是邻居。

对于有监督训练可以使用每个节点的预测label和真实label的交叉熵作为损失函数。

# 3 实验

实验主要分为三个部分：

（1）利用Web of Science引文数据集对学术论文进行分类；

（2）将贴子分类到不同的社区；

（3）利用生物蛋白之间的相互作用PPI对蛋白质功能进行分类。

实验过程中，所有预测节点的信息在训练过程中是不会出现的。实验设置如下，设置四个baseline，分别为

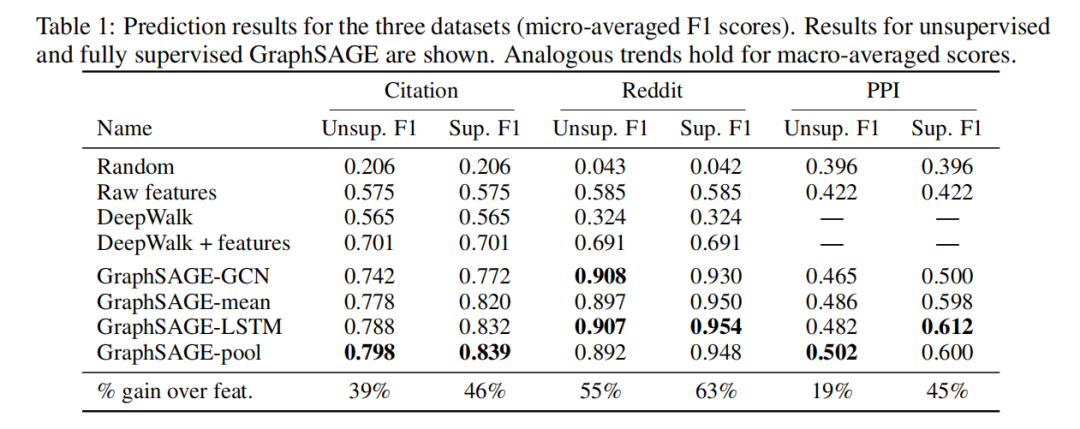
（1）随机分类器

（2）逻辑回归分类器

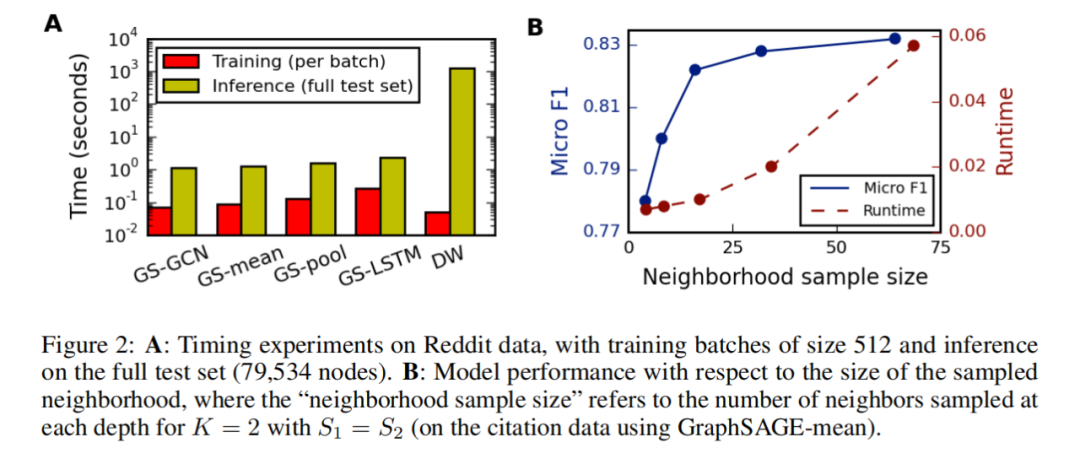
（3）DeepWalk算法

（4）原始特征+DeepWalk嵌入表达结果

同时对比四种不同的GraphSAGE算法（利用不同的聚合函数），对于GCN版本，无监督变体采用上文的损失函数，有监督变体利用分类交叉熵损失。非线性函数均采用ReLU函数，K=2，S1=25，S2=10。实验结果如下：



下图A是训练和测试时间的实验结果，B是采样邻域大小对性能影响的结果。



# 4 实践

参考github：

<https://github.com/jpegbert/NLP_Coding/tree/master/graphsage>

# 5总结

GraphSAGE的精华在于inductive learning，inductive learning可以在测试时对新加入的节点进行推理，这在实际场景的应用中是非常重要的。

GraphSAGE采用了采样的机制，客服了GCN训练时内存和显存上的限制，使得图模型可以应用到大规模图结构中，是目前几乎所有工业上图模型的雏形。

参考：

<https://mp.weixin.qq.com/s/pn4GYn9CU-sq49qg2OHkIw>

<https://mp.weixin.qq.com/s/4XhOp8AQUfebLCeAK0xVcQ>

<https://mp.weixin.qq.com/s/1DHvLLysMU24dBeLzbSpUA>

<https://mp.weixin.qq.com/s/6uwg2-ORFEv9hY7utYs_ow>

<https://zhuanlan.zhihu.com/p/79637787>

<https://mp.weixin.qq.com/s/4Si-I8NN1rDAJVCN7PS88g>

<https://mp.weixin.qq.com/s/QcLTaNlROz1x4zx70XYtrg>